TO WHAT EXTENT ARE FORMATIVE ASSESSMENT STRATEGIES USED IN SCHOOLS CONTRIBUTING TO STUDENT LEARNING? A SYSTEMATIC REVIEW AND META-ANALYSIS

DEAN A. DUDLEY

from

UNIVERSITY OF OXFORD

DEPARTMENT OF EDUCATION

2023

Thesis submitted in partial fulfilment of the requirements

for the award of the degree

MASTER OF SCIENCE (EDUCATIONAL ASSESSMENT)

Deposit and Consultation of Thesis

One copy of your dissertation will be deposited in the Department of Education Library via Oxford Research Archives where it is intended to be available for consultation by all Library users. In order to facilitate this, the following form should be inserted in the library copy of the dissertation.

Note that some graphs/tables may be removed in order to comply with copyright restrictions.

Surname	Dudley
First Name	Dean Alan
Faculty Board	Education
Title of Dissertation	To what extent are formative assessment strategies used in schools contributing to student learning? A systematic review and meta-analysis.

Declaration by the candidate as author of the dissertation

1. I understand that I am the owner of this dissertation and that the copyright rests with me unless I specifically transfer it to another person.

2. I understand that the Department requires that I shall deposit a copy of my dissertation in the Department of Education Library via Oxford Research Archives where it shall be available for consultation, and that reproductions of it may be made for other Libraries so

that it can be available to those who to consult it elsewhere. I understand that the Library, before allowing my dissertation to be consulted either in the original or in reproduced form, will require each person wishing to consult it to sign a declaration that he or she recognises that the copyright of this thesis belongs to me. I permit limited copying of my dissertation by individuals (no more than 5% or one chapter) for personal research use. No quotation from it and no information derived from it may be published without my prior written consent and I undertake to supply a current address to the Library so this consent can be sought.

3. I agree that my dissertation shall be available for consultation in accordance with paragraph 2 above.

Declaration

I, Dean A. Dudley, declare that this thesis, submitted in partial fulfilment of the requirements for the award of Master of Science (Educational Assessment), in the Department of Education, University of Oxford, is wholly my own work unless otherwise referenced or acknowledged. Research assistants were used in the double-blinded screening processes only to maintain methodological rigor. They did not contribute to the literature review, analysis, or writing of this dissertation in anyway. The document has not been submitted for qualification at any other academic institution.

Dean A. Dudley, CF
PhD, M.Ed (Research), CAEL, Grad Dip Ed (Secondary), B.HSc

Dedication

This dissertation is a heartfelt tribute to my family. Despite my parents not completing high school, they instilled in me the importance of education from an early age. My mother, Alison, and my father, Alan, have been incredibly supportive and have always been the foundation of my unwavering commitment to learning. Their belief in my abilities has shaped who I am today.

I am also grateful to my two younger brothers, Ryan and Trent Dudley. Despite our contrasting personalities, we firmly believe in the power of family bonds. Whether it was on the football field, in the schoolyard, or in our professional lives, we have always put family first.

Above all, I dedicate this thesis to the three remarkable women who stood by me throughout my apparent never-ending journey of scholarship. My wife, Ana, and my two daughters, Carolina and Georgia, are the very essence of my existence. I want them to know that without their enduring love and unwavering support, everything else would lack purpose and significance.

Acknowledgements

I am grateful to my primary supervisor, Dr Stuart Cadwallader who provided guidance and support in all facets of this project. Thank you, Stuart, for treating me as a colleague throughout my studies here at the University of Oxford.

My collaborators and research assistants on this study, Dr Erin Mackenzie, Associate Professor Stuart Woodcock, Dr Melissa Johnstone, Elisabeth Hitchens, and Hayley Dean who shared the high and lows of this study from inception to completion. I am thankful for the countless hours they spent on the data trail with me. I consider them the most professional teachers I have ever worked with, and I look forward to much more collaboration in the coming years.

I also need to thank my friends and colleagues who shared my Oxford journey. Specifically, I would like to mention Rebecca Dowbiggin, Bud Teasley, and Krystal Ketcher for their support and sharing this experience with me.

Finally, it would be remiss of me not to thank the fantastic staff I work with at Macquarie University, University of Queensland, University of Sydney, and the University of the South Pacific who have all supported me professionally to complete yet another degree.

Abstract

Objective: To determine the effects of formative assessment interventions and strategies aimed at optimising the learning outcomes in children and adolescents.

Design: A systematic review and meta-analysis.

Data sources: After searching PsycInfo, ERIC, Education Research Complete, A+ Education and Scopus electronic university databases, there were 3689 potentially eligible studies published between January 1, 2002 to December 31, 2022.

Eligibility criteria for selecting studies: This study includes randomised controlled trials, quasi-experimental studies, and controlled trials that assessed the effect of a formative assessment-based strategy or intervention against learning outcomes in students attending either primary or secondary years of schooling (i.e., aged 5–18 years).

Results: Thirty-eight (38) studies with over 16,000 participants and 43 calculated effect sizes were included in the final analysis. The sample size adjusted mean effect across all the learning and development outcomes was of a small to medium magnitude (Hedges' *g*=0.31), 95% confidence interval [CI] (0.29 to 0.39). After applying the Duval and Tweedie (2000) Trim and Fill Method to account for publication bias, the average effect size was adjusted to g=0.16. Effect sizes also varied significantly based on whether they targeted cognitive, psychomotor, or affective learning outcomes. Only a small proportion (less than 10%) of formative assessment interventions examined in this systematic review indicated a detrimental impact on student learning and development.

Conclusion: Formative assessment interventions consistently have a positive effect on student learning in both primary and secondary school settings across psychomotor, affective, and cognitive learning domains. However, the reported effect of formative

assessment interventions in previous meta-analytic studies appears to have been inflated by the inclusion of studies that suffer from weaker study designs. Furthermore, the scope of studies included in this systematic review and meta-analysis showed even studies with stronger study designs were plagued with poor methodological quality and substantial publication bias. Moreover, the evidence reported is limited by consistency in intervention dosage, study design, and data collection instruments.

Table of Contents

List of Tables

List of Figures

List of Acronyms and Abbreviations

CT – Controlled Trial

ES – Effect Size

OECD – Organization for Economic Cooperation and Development

QE – Quasi Experimental Study

RCT – Randomised Controlled Trial

CHAPTER 1

Introduction

1.1 Background

The terms "formative assessment" (Bloom, 1968) and "assessment for learning" (Mittler, 1973) have existed in the educational literature for over 50 years, however, widespread utilisation of these terms in teaching practice has only emerged in the past few decades. While there is ongoing debate regarding consistent terminology in this field, recent literature indicates the presence of a substantial body of studies sharing a common purpose of formative assessment. These studies span multiple domains of learning and can be effectively combined for meta-analytic synthesis (Ozan & Kincal, 2018).

According to Moss and Brookhart (2019), formative assessment is a process of gathering information about students' learning progress during the instructional process. It aims to provide teachers with feedback about their students' understanding of concepts and skills in real-time, enabling them to adjust their teaching to meet students' needs. Formative assessment is often contrasted with summative assessment, which evaluates student learning at the end of a unit or course.

Formative assessment can take many forms, including observation, questioning, self-assessment, peer-assessment, and teacher feedback (Wylie & Lyon, 2016). For example, teachers may use student-generated questions, exit tickets (i.e., tests of knowledge before leaving a class), or quick quizzes to gather information about students'

understanding of a topic and identify areas for further instruction. Alternatively, teachers may use self-assessment tools that encourage students to reflect on their learning and provide feedback to their peers (Wylie & Lyon, 2016).

Traditionally, formative assessment was seen as a way for teachers to monitor student progress, identify gaps in knowledge, and adjust their teaching to meet students' needs (Heritage, 2007). This often involved administering tests and quizzes and grading them to assess students' understanding of a particular topic. While this approach provided some valuable feedback to teachers, it often failed to engage students actively in their own learning and did not provide them with an opportunity to self-reflect and improve their own learning (Brookhart, 2011).

In recent times, the notion of formative assessment has undergone a transformation, placing students at the forefront of their own learning journey. These contemporary interpretations of formative assessment encompass a variety of assessment strategies that empower students to actively monitor their progress, recognise their strengths and weaknesses, and take appropriate steps to enhance their learning (Forrest, 2018). Additionally, they involve leveraging technology to gather and analyse data pertaining to students' learning progress (Bhagat & Spector, 2017; Elmahdi, Al-Hattami, & Fawzi, 2018). This may include the utilisation of online assessment tools that offer teachers real-time feedback on students' grasp of specific topics. Such tools often incorporate dashboards and visualisations, enabling teachers to promptly identify areas of proficiency and areas that require attention, thereby enabling them to tailor their instruction to meet the needs of their students (West, 2012).

Formative assessment continues to be extensively researched and discussed in the literature for the numerous advantages it affords for both students and teachers. For students, it has been long argued formative assessment plays a decisive role in enhancing learning outcomes by enabling them to pinpoint specific areas requiring focused attention and effort (Black & Wiliam, 1998; Hattie & Timperley, 2007). It is also believed that formative assessment may foster a growth mindset by instilling the belief that mistakes are valuable opportunities for learning and growth (Dweck, 2006). Moreover, some researchers claim that formative assessment actively engages students in the learning process and therefore enhances student engagement and motivation (Heritage, 2010).

For teachers, seminal formative assessment research argues that it serves as a powerful tool for identifying areas where students may be struggling, allowing for timely interventions and adjustments in instruction to address their needs (Sadler, 1989; Wiliam, 2011). Furthermore, it is positioned as a means that provides invaluable feedback on the effectiveness of teaching strategies that enables teachers to reflect on their practices, make necessary modifications, cultivate teacher-student relationships, afford opportunities for personalised feedback, foster individual growth, and build trust (Popham, 2008).

Formative assessment, despite its reported benefits, poses several pedagogical and implementation challenges. One significant challenge is the substantial time and effort demanded to effectively execute formative assessment practices. This encompasses designing and administering assessments, analysing the gathered data, and subsequently adapting teaching strategies accordingly (Heritage, 2007). Moreover, successful

implementation of formative assessment often necessitates extensive professional development to ensure that teachers possess the requisite skills and knowledge to utilise assessment strategies effectively (Guskey, 2007; Darling-Hammond & McLaughlin, 2011).

Whilst formative assessment has now become part of contemporary educational thought and vernacular, it is Dylan Wiliam and Paul Black who are credited with its popularity. In their definition, they describe formative assessment as the set of activities undertaken by teachers and/or students to provide feedback and modify teaching and learning activities (Black & Wiliam, 1998b, pp. 7–8). This definition offers multiple areas of focus for conducting formative assessment and to some extent explains the evolution in the concept in the last 25 years.

In a paper published later that year, Black and Wiliam (1998a) sought to determine the efficacy of formative assessment in educational practice. They conducted a detailed synthesis of the literature on formative assessment that has been widely cited and is still considered a seminal paper on the subject. They pointed out that there is a certain level of cohesiveness within the formative assessment literature, but the underlying differences between studies are such that amalgamating their results would not be meaningful. That said, the authors themselves have stated that they did not use any quantitative meta-analytic techniques on the data they gathered to support their claims statistically.

Black and Wiliam (1998a) also pointed out that the pedagogy associated with each study in their synthesis paper is likely to be very different from one study to the next, and that researchers tend not to collect or report the same type of information about their

studies, leading to a lack of conceptualisation of the issues involved with formative assessment.

Notwithstanding these constraints, the work of Black and Wiliam (1998a) stands out as the most frequently referenced evidence supporting the favourable influence of formative assessment practices on student achievement, with over 3000 CrossRef citations at the time of writing. Black and Wiliam (1998a) specifically reported that formative assessment experiments yielded typical effect sizes ranging from .4 to .7 of a standard deviation, showcasing substantial improvements in student learning. Moreover, they noted that these effect sizes exceeded those commonly observed for other educational interventions (Black & Wiliam, 1998a).

Since the original review conducted by Black and Wiliam (1998a) on formative assessment, researchers have frequently referenced their reported effect sizes as evidence supporting the use of formative assessment in educational settings. For instance, Yeh (2010) includes Black and Wiliam (1998a), among other studies, to support the assertion that the effect size for testing feedback is at least .7 of a standard deviation in student learning.

A commentary piece featured in a leading education journal emphasised the significant influence of Black and Wiliam's seminal review as the most frequently cited reference on formative assessment, contributing to the widely accepted notion that "Everyone knows that formative assessment improves learning" (Shepard, 2005, p. 8). The authors further asserted that formative assessment surpasses other typical educational

interventions, yielding effect sizes ranging from .4 to .7, signifying its profound impact on learning outcomes (Shepard, 2005).

Similarly, in a study published in the very same year, it was contended that formative assessment, as supported by Black and Wiliam's work and when utilised effectively, stands as one of the most potent tools available for informing classroom decisions (Dorn, 2010).

Black and Wiliam's (1998a) review of formative assessment, while highly influential, has faced scrutiny regarding the magnitude of the effects attributed to it. Subsequent researchers, such as Dunn and Mulvenon (2009), have raised concerns about methodological issues (i.e., confounding treatments, non-equivalent groups) and limitations present in the studies included in Black and Wiliam's review.

Another study used to support Black and Wiliam's claims, conducted by Martinez and Martinez (1992), examined the effects of assessment frequency on student achievement but did not adequately consider key elements associated with formative assessment, such as feedback and instructional differentiation. Critiques of Black and Wiliam's review suggest that the studies cited may not actually provide sufficient support for the claimed effects of formative assessment. Rather, they argue that instead of making definitive claims, Black and Wiliam should have highlighted the need for further empirical research in the field of formative assessment at the time (Dunn & Mulvenon, 2009).

Although these critical reviews of the seminal Black and Wiliam (1998a) paper offer valuable insights into the literature, they do not constitute the insight that can be drawn through meta-analyses. In other words, whilst they all agree that the processes of formative assessment may provide worthwhile learning effects in students, they cannot speak to the magnitude of that effect. Furthermore, it is important to note that critical reviews rely on reviewer judgment and are challenging, if not impossible, to replicate. Finally, the disparate conclusions being drawn in these critical reviews occur despite examining the same set of studies, emphasising the complexities and subjectivity inherent in evaluating the efficacy of formative assessment from a strictly qualitative perspective.

1.2 Previous Meta-Analyses of Formative Assessment

Recognising the limitations in the initial meta-analytic work conducted on formative assessment since the Black and Wiliam (1998a) review, Kingston and Nash (2011) conducted their own meta-analytic study that combined the results of studies that assessed the impact of using formative assessment practices in K-12 (i.e., students aged 5-18) classrooms (independent variable) on student achievement (dependent variable) in an effort to identify strategies that may help maximise the effects of the formative assessment process. The final sample consisted of 13 studies with 42 independent effect sizes. While the formative practices varied greatly across the studies analysed, they all have one common defining feature, which is that information was gathered and used with the intent of assisting in the learning and teaching process.

Of the 42 effect sizes identified by Kingston and Nash (2011), 19 were based on mathematics formative assessment, 12 on reading, language arts, or writing, 10 on science, and one on music. With respect to grade level categories of students, two effect sizes involved students in Kindergarten through the 4th grade (i.e, students aged between approximately 5-10 years), 16 involved 5th through 7th graders (i.e, students aged between approximately 10-14 years), 16 involved 8th through 9th graders (i.e, students aged between approximately 14-16 years), and eight involved 10th through 12th graders (i.e, students aged between approximately 16-18 years).

Interestingly, the Kingston and Nash (2011) meta-analysis managed to identify five approaches to formative assessment that were emanating from the literature. They were

1. *Professional development* which was defined as involving educators spending a period of time learning and focusing on how to implement various aspects of formative assessment techniques (e.g., comment-only marking, self-assessment, etc.) into their respective classroom instruction.
2. *Curriculum-embedded assessment* which required teachers to manage open-ended formative assessments at critical points throughout the curriculum instruction process so that they able to gain an understanding of the students' learning progression.
3. *Use of a computer-based formative assessment* which involved the online management of indicator level tests that in turn provided progress reports to teachers. They may have also incorporated a tutoring feature in the form of student-level learning scaffold, assessment rubric, or learning sequence.

4. *Use of feedback* which required teachers to provide students with detailed feedback about their performance prior to any summative assessment as a guide for them to understand their learning processes.
5. Finally, various *other types of formative assessment* were recognised. These encompassed classroom assessments, designed to furnish teachers with insights into student progress and inform adjustments to their teaching approaches. Furthermore, students may have also been encouraged to engage in reflection to evaluate their learning journey, and/or verbal assessment techniques were integrated into ongoing classroom activities through assessment conversations.

Whilst identification of these categories of formative assessment interventions are useful and provide a potential filter for future reviews, they are limited by the dependant variable being student achievement and thereby limiting the scope of learning and development that formative assessment strategies may be influencing. In their recent meta-analysis of physical education interventions, Dudley et al (2022) synthesised learning and development interventions deriving from up to four learning domains (i.e., cognitive, affective, social, and psychomotor). It therefore may be worthwhile considering whether formative assessment interventions have been utilised outside of developing merely cognitive learning domain outcomes.

Furthermore, Kingston and Nash's (2011) calculation and analysis of effect sizes require greater attention as meta-analytic techniques have advanced substantially in the last decade (Borenstein et al., 2021). First, the effect size used to quantify the magnitude of the relationship between the independent variable and the dependant variable in the Kingston

and Nash (2011) study were based on Cohen's *d*. However, given that none of the 13 included studies had samples that contained more than 23 participants (six studies had participants numbered in single digits), other indices of standardised difference might have been considered. Case in point, Hedges (1981, p. 111) showed that Cohen's *d* estimates are positively biased by small studies and proposed that unbiased standardised effect estimates were needed for meta-analytic studies with small sample sizes. Compared to Cohen's d, Hedges' *g* estimates the standardised difference between the means of two populations by providing a bias correction (using the exact method) to Cohen's *d* for small sample sizes (<20 participants). That said, effect size calculations are roughly equivalent for both *d* and *g* when the analysis incorporates more than 20 participants (Hedges, 1981).

Another evolution in analysis of effect sizes that was not reported by Kingston and Nash (2011) was a full suite of tests of heterogeneity. Tests of heterogeneity determine the variability among the included studies in terms of their outcomes. It is a critical consideration in meta-analytical studies as it can significantly impact the validity and generalisability of the findings.

When conducting a meta-analysis, the primary objective is to synthesise the results from multiple individual studies to derive an overall estimate of the effect of the phenomena under investigation. However, the presence of substantial heterogeneity among the included studies poses challenges in accurately interpreting any estimate of effect. A high degree of heterogeneity indicates substantial differences in the studied populations, interventions, study designs, and/or measurement methods.

Borenstein et al (2021) argue that reporting a comprehensive suite of heterogeneity tests is crucial to ensure the reliability and rigor of the conclusions drawn from meta-analyses are maintained. Kingston and Nash (2011) however only reported two tests of heterogeneity which were the Q-statistic (Cochran, 1954) and the I^2 statistic (Higgins & Thompson, 2002). Borenstein et al (2021) and Rücker et al (2008) recommend that in addition to both Q and I^2 statistics, that at least a Tau statistic (T^2) and a prediction interval are also included in the tests of heterogeneity.

Finally, several included studies were derived from unpublished Masters or Doctoral theses, there was no reporting of where study designs show risk of bias (i.e., quality of study methodology), and no statistical calculation of publication bias (neither a selection model nor a funnel-plot-based method) that might exist in the universe of studies included in their analysis. According to Lin and Chu (2017), publication bias is a serious problem in meta-analyses, which can affect the validity and generalisation of conclusions being drawn and is therefore important to report methodological factors that expose risk of bias and also quantify publication bias using measures that permit comparisons between different meta-analyses.

In 2015, Graham and colleagues conducted a meta-analysis of formative assessment practices and writing performance (Graham et al., 2015). They sought to determine whether formative writing assessments that were directly tied to everyday classroom teaching could enhance writing performance in students between grades 1-8. The strength of their meta-analytic technique was they only included true and quasi-experimental studies in their

analysis however nine of these studies were located in unpublished Masters or Doctoral dissertations rather than in peer-reviewed journal articles.

Graham et al. (2015) employed a methodology that involved calculating an average weighted effect size when there was a minimum of four independent comparisons assessing the effectiveness of formative assessment strategies. The findings of their study revealed a statistically significant improvement in writing performance through formative writing assessments. However, the adjusted weighted effect sizes varied across different feedback sources, with a higher effect size of d^{adj}=0.87 observed when feedback was provided by adults, while a lower effect size of d^{adj}=0.38 was observed when feedback was received from a computer.

Like Kingston and Nash (2011), Graham et al (2015) report limited measures of heterogeneity in their meta-analyses (Q and I^2 statistics only) and there is no reporting of risk of bias nor an analysis of any publication bias.

In 2017, the Institute of Education Sciences within the U.S. Department of Education commissioned Klute and colleagues to review the evidence of formative assessment and elementary school student academic achievement (Klute et al., 2017). Of the 30 separate effect sizes they calculated, the mean effect size of formative assessment on student achievement was just over a quarter of a standard deviation (d=0.26). The average effect size for formative assessment in mathematics was reported as d=0.36 but only d=0.22 for reading and d=0.21 for writing. The range of effect sizes was also substantive. They ranged from d=-0.46 to d=1.22 (Klute et al., 2017).

The meta-analytic limitations of the Klute et al (2017) study include the absence of any statistical test of heterogeneity. Heterogeneity is only discussed in broad qualitative terms throughout the report. Secondly, like earlier reviews on this topic, there is an absence of risk of bias reporting and statistical calculations of publication bias.

Most recently, Lee and colleagues (2020) published a systematic review and meta-analysis on the effectiveness and features of formative assessment in the United States education system between grades K-12. They studied the features of formative assessment interventions and their impacts on student learning. In their analyses of 33 studies which all included a control condition, there was an overall small positive effect of formative assessment on student learning (d=0.29). Unlike previous systematic reviews and meta-analyses, Lee et al (2020) conducted further investigations using meta-regression analyses. By doing so, they were able to report that supporting student-initiated self-assessment (d=0.61) and providing formal formative assessment evidence (e.g., written feedback on quizzes; d=0.40) via a medium-cycle length (within or between instructional units; d=0.52) improved the effectiveness of formative assessments.

This most recent meta-analysis by Lee et al (2020) instigated several strengths to their design that had not occurred in the three earlier meta-analytic studies. The first was only including studies that had a control/comparison group. This substantially increased the causal generalisability of their findings. The second was the inclusion of a meta-regression analysis that they used to determine the factors within the formative assessment interventions that were contributing to the changes in the observed variance. They also calculated their effect sizes using Hedge's g (Hedges, 1981) to adjust for the upward bias of

including studies with a small sample size. Finally, they performed two separate statistical tests of publication bias (i.e., Egger's Intercept test - Egger, Smith, Schneider, & Minder, 1997; and Trim and Fill method - Duval & Tweedie, 2000) from their universe of studies.

However, the Lee et al (2020) meta-analysis was limited by not reporting any heterogeneity analysis of the included studies. Furthermore, studies were limited to the context of being implemented in the United States and 14 of the 33 studies included in the analysis were derived by unpublished Masters or Doctoral dissertations rather than peer-reviewed journal articles.

It is important at this point to acknowledge that while meta-analyses always possess limitations, they can effectively supplement qualitative judgment-based reviews. There are significant differences present across the spectrum of studies initially reviewed by Black and Wiliam (1998a) and it is likely that these differences extend to any set of studies related to the examination of formative assessment. By limiting the inclusion criteria in future meta-analyses, it is possible to focus on a more homogenous set of studies, which may in turn yield outcomes with greater insight.

Additionally, Black and Wiliam's (1998a) argument that pedagogical differences across studies reduce the feasibility of conducting a meta-analysis is a challenge faced by any meta-analyst (Glass, 2000). That is, mean effect sizes produced in a meta-analysis may not be meaningful if aggregated over studies with disparate independent or dependent variables (Lipsey & Wilson, 2001). Nevertheless, contemporary meta-analytic techniques can provide evidence regarding the compatibility of results and whether combining them

would result in meaningful outcomes. In other words, contemporary meta-analyses aim not only to determine a grand average effect size but also to identify potential sources of variability in study findings.

Finally, all meta-analyses of formative assessment in the literature to date have focussed exclusively on learning derived from the cognitive learning domain. Many scholars now argue that a holistic and integrated understanding of learning is warranted to understand human development whereby psychomotor and affective learning are also considered (Baxter Magolda, 2000; Kolb & Kolb, 2009).

For clarity, psychomotor, affective, and cognitive learning are distinct domains that encompass different aspects of a student's learning process. Understanding these domains facilitates an understanding of the multifaceted nature of any learning outcomes being measured.

Cognitive learning primarily focuses on intellectual abilities, knowledge acquisition, and higher-order thinking skills. It involves the development and application of cognitive processes such as comprehension, analysis, synthesis, and evaluation (Hoque, 2016). Cognitive learning is often associated with the acquisition of information, problem-solving abilities, critical thinking, and conceptual understanding.

On the other hand, affective learning refers to the development of attitudes, values, beliefs, and emotions that influence learners' motivation, interests, and behaviours. It encompasses the emotional and social aspects of learning, including self-awareness,

empathy, interpersonal skills, and ethical values (Hoque, 2016). Affective learning aims to foster positive attitudes, ethical decision-making, and social-emotional development.

Lastly, psychomotor learning pertains to the development of physical skills, coordination, and motor abilities. It involves the acquisition and refinement of motor skills through practice and physical performance, such as kicking a football, playing a musical instrument, or even handwriting (Hoque, 2016). Psychomotor learning emphasises hands-on experiences, muscle memory, and the integration of sensory and motor processes.

By recognising these three domains of learning in the educational literature, educators and researchers can gain a deeper appreciation of the true effect and potential of formative assessment strategies employed in their teaching practices.

1.3 Research Questions

As such, the research questions in the present analysis are as follows:

1. What is the sample size adjusted effect size, heterogeneity, and publication bias evident of formative assessment interventions on learning and development in primary and secondary school settings?
2. To what extent is the reported effect size, heterogeneity, and publication bias of formative assessment on learning and development constrained by learning domain?
3. What extent is the reported effect size, heterogeneity, and publication bias on learning and development limited by specific formative assessment practices in primary and secondary school settings?

CHAPTER 2

Methods

2.1 Protocol and Registration

To enhance the reporting rigor, the PRISMA 2020 Checklist (Page et al., 2021) was adhered to as closely as possible considering that this systematic review and meta-analysis focused on educational literature rather than clinical or medical science. Furthermore, since the study did not primarily examine health-related outcomes, it was not eligible for registration with the International Prospective Register of Systematic Reviews.

Covidence(TM) software, as reviewed by Babineau (2014), was used in the title and abstract screening, full-text review, methodological quality, and data extraction phases of the study. Covidence(TM) is an online software tool that manages and organises the process of creating systematic reviews.

2.2 Eligibility Criteria and Study Selection

To be eligible for inclusion in the review, studies needed to meet the following criteria: (1) Participants: school-aged children and adolescents aged 5–18 years old that were enrolled in school. No children were excluded regardless of specific learning needs; (2) Intervention characteristics included studies that used formative assessment(s) at school as the intervention medium; (3) Comparison: control group that received the regular instruction or no formative assessment; (4) Outcomes: psychomotor (e.g., gross motor skill, motor competence, fundamental movement skill acquisition), cognitive (e.g., executive

function, memory, attention, academic scores); or affective (e.g., motivation, self-esteem, self-efficacy, enjoyment, self-regulation) outcomes; (5) Study design: randomised controlled trials (RCTs), quasi experimental trials (QEs), and controlled trials (CTs).

2.3 Search Strategy

The search strategy entailed explorations of the A+ Education, Education Research Complete, ERIC, PsycInfo, and Scopus electronic databases. To ensure a comprehensive overview of the field, the search involved studies published for the last 20 years between 1st January 2002 and 31st December 2022. The search itself took place over five days between 3rd January to 8th January 2023. Only peer-reviewed journal articles published in English were included.

To identify relevant research, a series of blocked search terms were constructed. The first group *(Block 1)* identified relevant participants. These included variations of "adolesc*" OR "child*" OR "youth" OR "teen*" WITH “school” respectively. The second block of search terms identified published studies that were pertinent to formative assessment *(Block 2)*. This included the terms “formative assessment” OR “assessment for learning” OR “feedback”. The last group of search terms identified research that implemented relevant study designs *(Block 3)* and included variations of intervention, “random* control trial” OR quasi-experiment* OR “control trial” OR “program” OR “comparison”. The blocked search terms were then merged (*Block 1* AND *Block 2* AND *Block 3*) to be applied simultaneously in the electronic databases. Given the outcomes of interest included three learning domains, each representing a diversity of learned skill and

aptitude, this study did not specify any outcome search terms. Rather, the use of the screening process was used to identify studies with appropriate learning outcomes residing within each of the learning domains.

To complement the database searches, a bi-directional screening of articles was conducted using previously published systematic reviews in formative assessment. Bi-directional screening is a method where a reviewer screens all references within an article and any articles that cited the article (Hinde & Spackman, 2015). This process aims to include relevant articles that may have been missed through traditional database searching.

2.4 Data Collection Process

All articles captured in the initial database search and complementary bi-directional screening were uploaded into the Covidence^{TM} review management software library. This software automatically removes any duplicate studies it detects. Titles and abstracts for the remaining studies were then reviewed by at least two independent reviewers from five different Australian universities (Macquarie University -DD, Griffith University - SW, University of Queensland –MJ; EH, Western Sydney University - EM, and the University of New South Wales - HD) for inclusion or exclusion into the systematic review, with any conflicts resolved using a third reviewer. Articles were excluded that were clearly unrelated to the topic, while those with the potential to be relevant were retained. This was followed by two reviewers independently screening the full-text articles for inclusion with any conflicts resolved using an independent third reviewer. Only those articles that met the study's inclusion criteria were retained. For each study, the following data was extracted:

(1) authors' names, publication year, and country of the study; (2) number of participants; (3) details about the formative assessment intervention, including pedagogical model, instruction model, and policy change; (4) information about the instruments used to measure psychomotor, cognitive, and affective social learning; (5) the targeted intervention site, such as primary or secondary school; and (6) results for all groups regarding the parameters of interest. Table 1 provides a summary of the included studies based on learning outcomes, study design, country, intervention mode, and intervention site. The meta-analysis of these data is presented through a series of forest plots based on the learning domain level of analysis (i.e., combined, cognitive, psychomotor, and affective learning effect).

2.5 Quality of Individual Studies

The methodological quality of the included articles after screening was completed using a 10-item quality assessment scale adapted from Van Sluijs et al. (2007) (see Table 1). Adapted versions of this tool have been used previously to report the quality of methodologies in educational research (Cotton et al., 2019; Dudley et al., 2015; Dudley et al, 2020; Dudley et al., 2022).

For each article included, two reviewers conducted independent assessments to determine the presence or absence of each assessed item. If an item was not adequately described, it was considered absent. In cases where there was not 100% agreement between the reviewers, a third reviewer evaluated the paper and made a final determination regarding the disputed item's presence or absence.

Table 1

Methodological quality assessment items (Adapted from van Sluijs et al., 2007)

Item	Description
A	Key baseline characteristics are presented separately for treatment groups and for randomised controlled trials and controlled trials, positive if baseline outcomes were statistically tested and results of tests were provided.
B	Randomisation procedure clearly and explicitly described and adequately carried out (generation of allocation sequence, allocation concealment and implementation)
C	Validated measures of learning (validation in same age group reported and/or cited)
D	Drop out reported and ≤20% for < 6-month follow-up or ≤30% for ≥6-month follow-up
E	Blinded outcome variable assessments
F	Learning assessed a minimum of 6 months after pre-test
G	Intention to treat analysis used (participants analysed in group they were originally allocated to, and participants not excluded from analyses because of noncompliance to treatment or because of some missing data)
H	Potential confounders accounted for in outcome analysis (e.g. baseline score, group/cluster, age)
I	Summary results for each group + treatment effect (difference between groups) + its precision (e.g. 95% confidence interval)
J	Power calculation reported, and the study was adequately powered to detect hypothesized relationships

2.6 Risk of Publication Bias

To determine the presence of publication bias in the studies, two statistical tests were implemented. The first test, Classic Fail-Safe N (Orwin, 1983), calculates the number of studies with an effect size of zero or less that would need to be added to the meta-analysis for the reported effect to lose statistical significance ($p<0.05$). The second test, Trim and Fill (Duval and Tweedie, 2000) is a three-step method that aims to identify and correct for funnel plot asymmetry resulting from publication bias.

Funnel plot asymmetry refers to an asymmetrical distribution of data points in a funnel plot, which is a graphic display commonly used in meta-analysis. In a funnel plot, the size of each study is plotted against its effect size. The plot usually exhibits a funnel-shaped pattern, where smaller studies are dispersed widely at the bottom and larger studies are clustered more closely towards the top (See Figure 3 as an example).

In a symmetrical funnel plot, the spread of points on both sides of the plot is roughly the same, and the plot is centred around an overall effect estimate. However, in an asymmetric funnel plot, the spread of points is uneven on one or both sides, suggesting that some studies are reporting effect sizes that differ from the overall estimate.

Asymmetry in a funnel plot may indicate the presence of publication bias, where studies with statistically significant results are more likely to be published and included in the meta-analysis, while those with non-significant results are less likely to be published or

included. Other sources of bias, such as selective outcome reporting or heterogeneity in study characteristics, can also contribute to funnel plot asymmetry.

To determine an adjusted effect size estimate if funnel plot asymmetry is evident initially involves the removal of smaller studies that cause funnel plot asymmetry. The second step involves the estimation of the true "centre" of the funnel plot using the trimmed data. In the final step, the trimmed studies and their missing "equivalents" are replaced around the centre. The adjusted intervention effect estimate is obtained by performing a meta-analysis that includes the filled studies. Additionally, the Trim and Fill method provides an estimate of the number of missing studies either side of the original mean estimate.

2.7 Data Synthesis and Analysis

The statistical analyses for this study were performed using Comprehensive Meta Analysis software (version 4.1). The random-effects model was employed according to the method outlined by DerSimonian and Laird (2015) to analyse the effects of the interventions. The effect sizes were expressed as standardized effect size (Hedges' g) to facilitate comparisons with other educational research studies whilst still adjusting for studies that had small sample sizes (Lin & Aloe, 2021). It is important to note that when multiple intervention groups utilising different strategies were included in a study, their data were analysed as independent studies. Furthermore, when a study reported results for more than one learning domain outcome, the data were analysed separately at the domain level.

Additionally, in cases where multiple tests assessing the same learning outcome variable were incorporated within a single study, a combined standardised effect size was computed (Borenstein, 2021). A random effects model was also employed to compare the differences in effect sizes between formative assessment interventions based on learning outcomes (e.g., cognitive, psychomotor, affective) and the pedagogical or intervention approach (e.g., peer assessment, computer-assisted feedback). When more than one age, grade, or class cohorts were incorporated into a study, their data were examined as combined samples. Finally, when multiple measurements were reported at follow-up, only the last measurement was considered in the analysis.

Heterogeneity was evaluated and reported across all studies, as well as at the learning domain level, using a range of complementary statistical analyses. These included the Q statistic, inconsistency index (I^2), Tau statistic (T^2), and prediction interval. The Q statistic tested the null hypothesis of whether all studies in the model shared a common effect size. The I^2 statistic indicated the proportion of observed variance attributed to actual changes in effect sizes rather than sampling error. The T^2 statistic determined the variance of true effect sizes, while the prediction interval provided a confidence interval (with 95% confidence limits) that encompassed the range of true effect sizes for all observed samples.

CHAPTER 3

Results

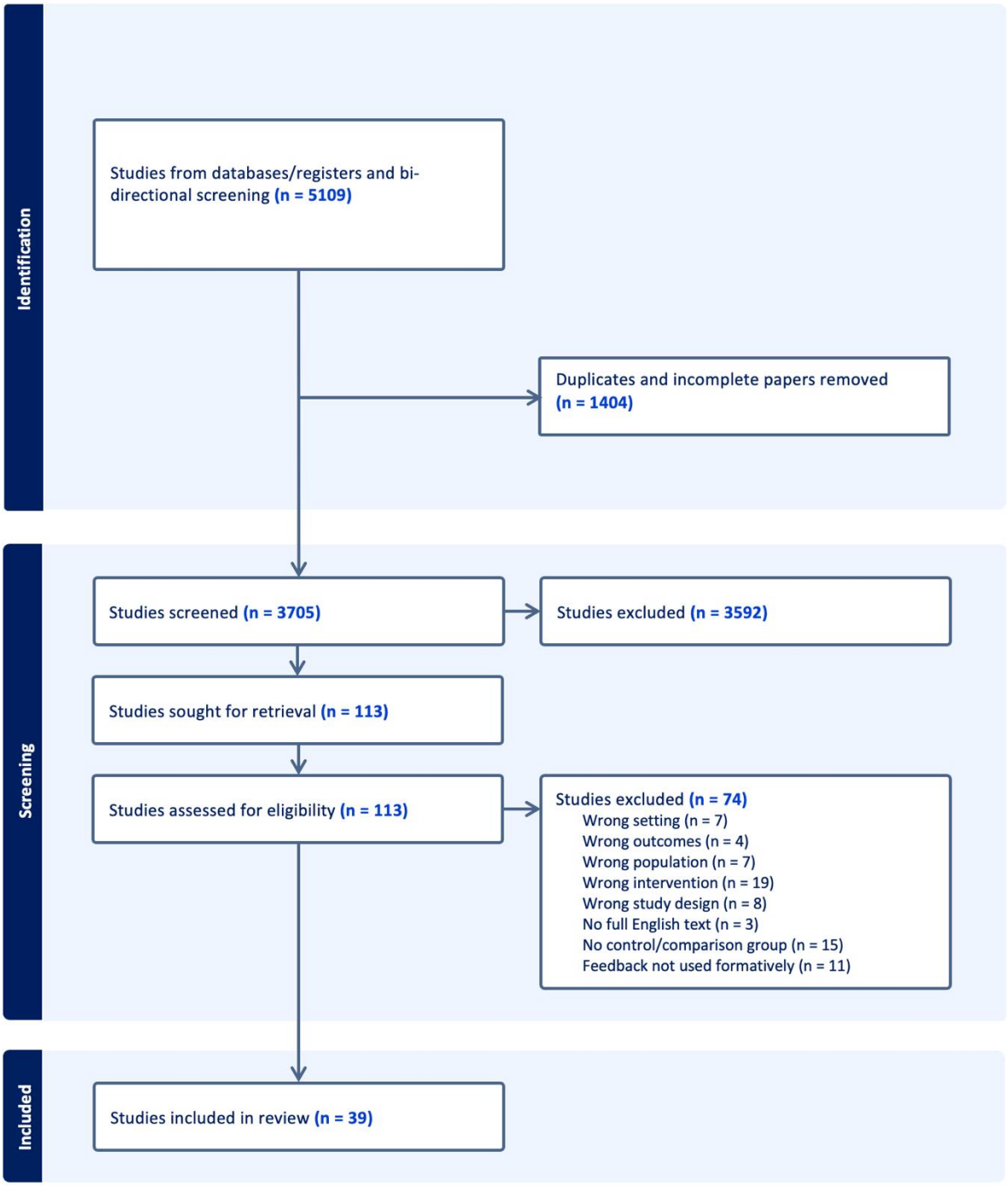
3.1. Study selection

The university database search strategy identified a total of 5,071 studies that were potentially eligible for inclusion. Through bidirectional screening, an additional 38 studies from four previously published reviews (Graham et al., 2015; Klute et al., 2017; Lee et al., 2020; Kingston & Nash, 2011) were added, resulting in a total of 5,109 studies for screening. After removing duplicate publications using Covidence^{TM} (n=1404), 3,705 papers underwent an assessment of their titles and abstracts. Out of these, 3,592 papers were deemed irrelevant by two reviewers based on their title and abstract, leaving 113 papers that underwent a full-text review.

At this stage, an additional exclusion criterion was added due to the number of title and abstracts revealing feedback was being used as an intervention strategy. Therefore, during the full-text review phase, interventions that used feedback were required to explicitly articulate how this was occurring formatively and not merely summatively. Consequently, an additional 74 studies were excluded by double-blinded consensus (as previously discussed) and their reasons for exclusion are detailed in the flowchart in Figure 1.

Figure 1

Flowchart of study selection



3.2. Study characteristics

There were 39 studies remaining from the study selection process that subjected to quality assessment and data extraction (See Table 2). Thirty (30) studies were conducted in primary/elementary schools and nine (9) studies were conducted in secondary schools. They consisted of 18 randomised controlled trials, 12 controlled trials, and nine (9) quasi-experimental studies. Across the three learning domains of interest, most studies targeted a cognitive learning or development outcome (n=29). The remaining studies targeted an affective (n=6) or psychomotor (n=5) learning or development outcome. It is important to note that one study targeted more than one learning domain which is why there is a disparity in the number of studies recorded.

Overall, the studies captured by the systematic review process came from 18 different countries and included over 16,000 school student participants.

Table 2

An overview of the studies extracted from papers and included in the systematic review/meta-analysis

Author (Year)	Country	Study Design	Sample Size	Primary Mode of Formative Assessment Intervention	Targeted Learning Domain(s)	Targeted Level of Schooling
Alitto et al (2016)	USA	RCT	57	Peer-assessment	Cognitive	Primary School
Andrade et al (2008)	USA	CT	116	Self-assessment	Cognitive	Primary School
Beekman et al (2021)	The Netherlands	RCT	974	Peer & self-assessment	Cognitive	Primary School
Berger et al (2019)	Australia	QE	211	Computer-based feedback	Affective	Secondary School
Bohan et al (2022)	Ireland	QE	38	Group contingency	Affective	Secondary School
Bonneton-Botté et al (2020)	France	CT	233	Computer-based feedback	Psychomotor	Primary School
Bottge et al (2021)	USA	CT	31	Computer-based feedback	Cognitive	Secondary School
Bush (2021)	USA	RCT	1052	Computer-based feedback	Cognitive	Primary School
Cáceres et al (2021)	Chile	QE	64	Scaffolded elaborated feedback	Cognitive	Primary School
Chan et al (2019)	Hong Kong	RCT	276	Assessment for Learning	Affective & Psychomotor	Primary School
Chen et al (2021)	Taiwan	RCT	55	Collaborative reading annotation	Cognitive	Primary School
Cohen et al (2012)	USA	RCT	97	Aligned developmental feedback	Psychomotor	Primary School
Damhuis et al (2015)	The Netherlands	CT	61	Repeated testing with feedback	Cognitive	Primary School
Damhuis et al (2016)	The Netherlands	QE	202	Individualised feedback	Cognitive	Primary School
Dao et al (2021)	Vietnam	CT	34	Peer assessment	Cognitive	Secondary School
Decristan et al (2015)	Germany	RCT	722	Assessment for Learning & Peer assessment	Cognitive	Primary School
Fidalgo et al (2015)	Spain	RCT	41	Peer assessment	Cognitive	Primary School
Förster et al (2018)	Germany	RCT	462	Learning progress assessment	Cognitive	Primary School
Fyfe & Rittle-Johnson (2016)	USA	RCT	49	Computer-based feedback	Cognitive	Primary School
Gaintza & Goikoetxea (2016)	Spain	CT	41	Immediate feedback & self-assessment	Cognitive	Primary School
Gamlem et al (2019)	Norway	CT	1165	Responsive pedagogy	Cognitive	Secondary School
Gedikli & Buldur (2022)	Turkey	QE	24	Self-assessment	Cognitive	Primary School
Gijlers et al (2013)	The Netherlands	RCT	29	Computer-based feedback	Cognitive	Primary School
Heir & Eckert (2014)	USA	RCT	103	Performance feedback	Cognitive	Primary School

Kegel & Bus (2012)	The Netherlands	RCT	152	Computer-based feedback	Cognitive	Primary School
Koedinger et al (2010)	USA	QE	1240	Computer-based feedback	Cognitive	Secondary School
Luquin & García Mayo (2021)	Spain	CT	38	Model corrective feedback	Cognitive	Primary School
Meyan & Greer (2010)	USA	CT	7795	Computer-based feedback	Cognitive	Primary School
Nord et al (2017)	Sweden	RCT	365	Immediate feedback	Psychomotor	Secondary School
Prinsen et al (2013)	The Netherlands	QE	189	Computer-supported collaboration (Peer assessment)	Cognitive	Primary School
Puklavec et al (2021)	Croatia	CT	38	Video & verbal feedback	Psychomotor	Primary School
Roll et al (2011)	Canada	CT	58	Immediate metacognitive feedback	Affective	Secondary School
Simmons et al (2015)	USA	RCT	156	Curriculum embedded measures	Cognitive	Primary School
Sukhram & Monda-Amaya (2017)	USA	RCT	60	Corrective feedback	Cognitive	Secondary School
Tavsanli et al (2021)	Turkey	QE	40	Graphic organisers	Cognitive	Primary School
Truckenmiller et al (2014)	USA	RCT	85	Performance feedback	Cognitive	Primary School
Vidal-Abarca et al (2014)	Spain	CT	25	Computer-based feedback	Cognitive	Primary School
Wong (2017)	Singapore	QE	146	Self-assessment	Affective	Primary School
Yin et al (2014)	USA	RCT	50	Embedded formative assessment	Cognitive	Primary School

Note: RCT=Randomised Controlled Trial; CT=Controlled Trial; QE=Quasi-experimental

3.3. Risk of bias based on quality of individual studies

The assessment of methodological quality, as presented in Table 3, indicated significant variation in reporting among the 39 studies included in the analysis. Only a fifth (20%) of the papers fulfilled five or more of the quality assessment criteria outlined by van Sluijs et al. (2007). Among the least frequently reported quality assessment criteria in these studies were appropriate power calculations (5%; n=2), learning assessments conducted at least six months after pre-tests (8%, n=3), and blinded assessments (10%; n=4). By contrast, the most frequently reported items were key baseline characteristics being reported and statistically tested (74%, n=29), a summary of key findings (difference between groups) with reported precision (51%, n=20) and the reporting of validated measurement instruments (49%, n=19) respectively.

Table 3

Results of methodological quality assessment

Paper No	Author (Year)	Methodological Quality Assessment Items										No. of criteria met
		A	B	C	D	E	F	G	H	I	J	
1	Alitto et al (2016)	✓	✓	✓					✓	✓	✓	6
2	Andrade et al (2008)	✓										1
3	Beekman et al (2021)	✓	✓	✓			✓					4
4	Berger et al (2019)	✓		✓						✓		3
5	Bohan et al (2022)	✓		✓								2
6	Bonneton-Botté et al (2020)	✓							✓			2
7	Bottge et al (2021)	✓										1
8	Bush (2021)					✓			✓			2
9	Cáceres et al (2021)	✓							✓			2
10	Chan et al (2019)	✓	✓	✓				✓	✓	✓	✓	7
11	Chen et al (2021)											0
12	Cohen et al (2012)	✓										1

13	Damhuis et al (2015)	✓		✓					✓	2
14	Damhuis et al (2016)	✓							✓	2
15	Dao et al (2021)	✓								1
16	Decristan et al (2015)			✓	✓			✓	✓	4
17	Fidalgo et al (2015)	✓						✓	✓	3
18	Förster et al (2018)	✓		✓		✓	✓	✓	✓	6
19	Fyfe & Rittle-Johnson (2016)				✓			✓	✓	3
20	Gaintza & Goikoetxea (2016)	✓		✓	✓				✓	4
21	Gamlem et al (2019)					✓		✓		2
22	Gedikli & Buldur (2022)			✓			✓		✓	3
23	Gijlers et al (2013)	✓								1
24	Hier & Eckert (2014)	✓		✓						2
25	Kegel & Bus (2012)	✓		✓	✓	✓		✓	✓	6
26	Koedinger et al (2010)				✓			✓		2
27	Luquin & García Mayo (2021)	✓					✓			2

28	Meyan & Greer (2010)							✓	✓			2
29	Nord et al (2017)	✓	✓		✓	✓					✓	5
30	Prinsen et al (2013)	✓			✓			✓	✓	✓		5
31	Puklavec et al (2021)			✓						✓		2
32	Roll et al (2011)									✓		1
33	Simmons et al (2015)	✓		✓				✓	✓	✓		5
34	Sukhram & Monda-Amaya (2017)	✓	✓	✓						✓		4
35	Tavsanli et al (2021)	✓		✓						✓		3
36	Truckenmiller et al (2014)	✓		✓	✓				✓	✓		5
37	Vidal-Abarca et al (2014)	✓	✓	✓								3
38	Wong (2017)	✓		✓				✓				3
39	Yin et al (2014)	✓			✓					✓		3
	% of criteria present	74% n=29	15% n=6	49% n=19	21% n=8	10% n=4	8% n=3	21% n=8	41% n=16	51% n=20	5% n=2	

3.4. Summary of evidence

Out of the 39 studies included a total of 44 effect sizes were analysed and focused on learning outcomes in the cognitive, affective, and/or psychomotor learning domains. In total, 24 distinct intervention strategies were identified across the included studies. However, one study (Dao et al., 2021) was removed from the subsequent meta-analyses due to the fact it reported an effect size that was greater than 0.6 of a standard deviation to the nearest reported effect size estimate of the next closest study and the reported standard error did not overlap. Hence, the final analyses were derived from a universe of 43 effect sizes from 38 studies.

3.4.1. Combined learning effects of formative assessment

As mentioned previously, all analyses conducted in this study focused on formative assessment interventions targeting specific learning variables among students aged 5-18 years attending primary and/or secondary schools, where the school itself served as the intervention site. In each study, students were assigned to either:

i) a formative assessment intervention (intervention condition), or

ii) ii) their regular curricula or a comparative pedagogy (control/comparison condition).

The effect size calculated for each outcome variable was the standardised mean difference (Hedges' g) between the intervention and control groups at the end of the intervention period or at a later follow-up point. The studies included in this analysis were selected from a defined universe of potential studies based on the previously established

inclusion/exclusion criteria. Therefore, a random-effects model was employed, and the conclusions drawn from the analysis pertain to that specific universe of studies.

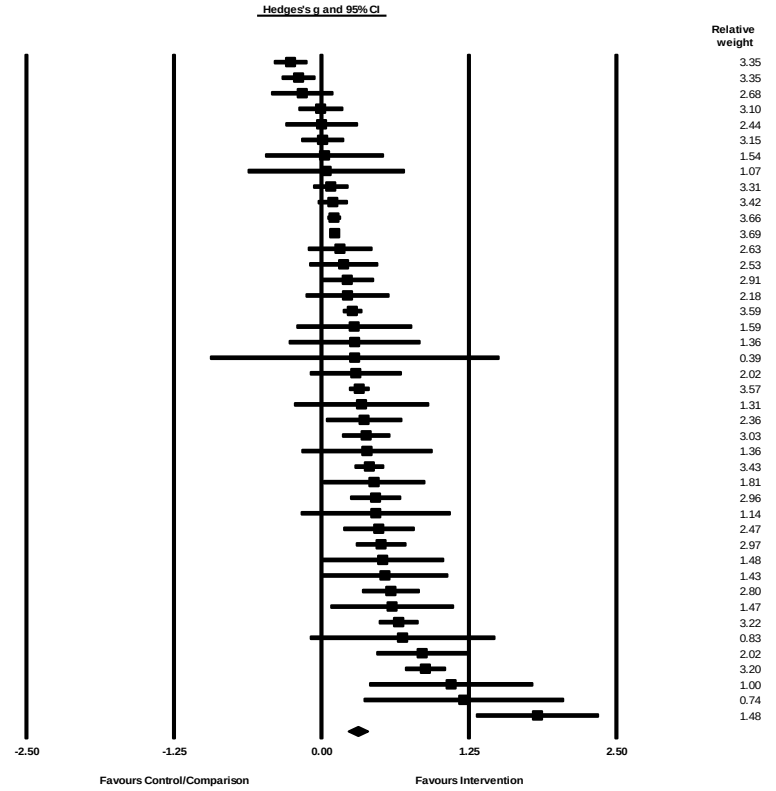
3.4.1.1 Do formative assessment interventions improve combined student learning outcomes?

When combined across learning domains and intervention medium, the formative assessment interventions included in this meta-analysis had a significant positive effect on student outcomes, $Z=7.492$, $p<0.001$. The standardised difference in means when adjusting for sample size was $g=0.31$ (95% CI [0.23, 0.39]), meaning that on average, students receiving a formative assessment intervention improved their learning or development by nearly a third of a standard deviation compared with those students who received their usual curriculum or pedagogy. Primary/elementary school interventions reported a slightly higher average effect size of $g=0.35$, whilst interventions conducted in secondary schools were substantially lower at $g=0.20$.

Figure 2

Combined effect of formative assessment by standardised difference in means (Hedges' g)

Study name	Subgroup within study	Outcome	Hedges's g	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value	Relative weight
Chan et al 2019b	Combined	Perceived competence & enjoyment	-0.261	0.071	0.005	-0.400	-0.122	-3.677	0.000	3.35
Forster et al 2018	Combined	Reading	-0.193	0.071	0.005	-0.332	-0.054	-2.724	0.006	3.35
Bonneton-Botte et al 2020	Handwriting test	Handwriting skills	-0.162	0.133	0.018	-0.423	0.098	-1.221	0.222	2.68
Decristan et al 2015b	Peer Assessment	Science Item Difficulty	-0.005	0.095	0.009	-0.191	0.181	-0.053	0.958	3.10
Fidalgo et al 2015	Combined	Writing quality	0.002	0.155	0.024	-0.302	0.306	0.014	0.989	2.44
Decristan et al 2015a	Formative Assessment	Science Item Difficulty	0.011	0.091	0.008	-0.167	0.188	0.119	0.905	3.15
Sukhram et al 2017	Fluency	Reading	0.027	0.255	0.065	-0.473	0.527	0.106	0.916	1.54
Bohan et al 2022	Combined	Academically engaged behaviour	0.041	0.338	0.114	-0.622	0.703	0.120	0.904	1.07
Koedinger et al 2010a	Regular	Mathematic achievement	0.080	0.075	0.006	-0.066	0.226	1.068	0.286	3.31
Gamlem et al 2019a	Norwegian National Test	Mathematic achievement	0.097	0.063	0.004	-0.027	0.221	1.536	0.124	3.42
Gamlem et al 2019b	Combined	Mathematic affective learning	0.108	0.027	0.001	0.055	0.161	3.985	0.000	3.66
Meyen et al 2010	Combined	Mathematic achievement	0.113	0.019	0.000	0.076	0.149	6.045	0.000	3.69
Berger et al 2019	CDSE	Decision Self Efficacy	0.158	0.137	0.019	-0.111	0.428	1.152	0.249	2.63
Prinsen et al 2013	Combined	Reading achievement	0.189	0.147	0.022	-0.099	0.478	1.287	0.198	2.53
Simmons et al 2015	Combined	Literacy	0.221	0.113	0.013	-0.001	0.442	1.950	0.051	2.91
Damhuis et al 2015	Combined	Vocabulary Acquisition	0.221	0.179	0.032	-0.130	0.573	1.233	0.217	2.18
Beekman et al 2021b	Combined	Self & External Regulation	0.264	0.040	0.002	0.185	0.343	6.553	0.000	3.59
Caceres et al 2021	Mathematic test	Mathematic achievement	0.278	0.248	0.062	-0.209	0.764	1.118	0.263	1.59
Fyfe et al 2016	Immediate feedback	Mathematic problem solving	0.282	0.283	0.080	-0.272	0.836	0.998	0.318	1.36
Bottge et al 2021	Mathematics test	Mathematic achievement	0.283	0.624	0.389	-0.940	1.505	0.453	0.650	0.39
Hier et al 2014	CBM-WE	Written expression	0.293	0.197	0.039	-0.093	0.678	1.487	0.137	2.02
Beekman et al 2021a	Combined	Self & External Regulation	0.321	0.043	0.002	0.237	0.405	7.504	0.000	3.57
Truckenmiller et al 2014	Correct writing sequence	Writing	0.340	0.290	0.084	-0.228	0.908	1.172	0.241	1.31
Kegel et al 2012	Tudor feedback	Literacy skills	0.362	0.163	0.026	0.043	0.681	2.223	0.026	2.36
Cohen et al 2012	Combined	Overhand Throwing	0.380	0.102	0.010	0.180	0.580	3.725	0.000	3.03
Yin et al 2014	Exam performance	Science achievement	0.387	0.281	0.079	-0.165	0.938	1.374	0.169	1.36
Bush 2021	Mathematics test	Mathematic achievement	0.406	0.062	0.004	0.284	0.528	6.528	0.000	3.43
Gaintza et al 2016	Combined	Spelling	0.447	0.219	0.048	0.017	0.877	2.036	0.042	1.81
Alitto et al 2016	Combined	Writing achievement	0.460	0.109	0.012	0.247	0.672	4.234	0.000	2.96
Luqui et al 2021	Model feedback	Secondary Language Acquisition	0.461	0.322	0.104	-0.171	1.093	1.430	0.153	1.14
Koedinger et al 2010b	Special Ed	Mathematic achievement	0.488	0.152	0.023	0.189	0.786	3.200	0.001	2.47
Nord et al 2017	Feedback	CPR performance	0.506	0.108	0.012	0.295	0.717	4.705	0.000	2.97
Roll et al 2011	Help condition	Self regulation	0.519	0.264	0.070	0.003	1.036	1.969	0.049	1.48
Chen et al 2021	Reading test	Reading comprehension	0.540	0.271	0.073	0.009	1.070	1.992	0.046	1.43
Chan et al 2019a	Overall FMS	Fundamental Movement Skills	0.589	0.123	0.015	0.348	0.830	4.792	0.000	2.80
Gijlers et al 2013	Combined	Science knowledge	0.600	0.265	0.070	0.080	1.120	2.263	0.024	1.47
Damhuis et al 2016	Combined	Vocabulary Acquisition	0.655	0.084	0.007	0.490	0.820	7.782	0.000	3.22
Vidal-Abarca et al 2014	PISA Test	Reading Literacy	0.688	0.399	0.159	-0.094	1.471	1.724	0.085	0.83
Andrade 2008	Essay scores	Writing achievement	0.855	0.197	0.039	0.469	1.240	4.345	0.000	2.02
Wong 2017	Combined	Perceived achievement	0.882	0.086	0.007	0.713	1.052	10.218	0.000	3.20
Puklavec et al 2021	Video Feedback all errors	Motor development	1.100	0.353	0.124	0.408	1.791	3.119	0.002	1.00
Gedikli et al 2022	MAIfC	Scientific metacognition	1.208	0.431	0.186	0.363	2.052	2.802	0.005	0.74
Tavsanli et al 2021	Combined	Writing	1.832	0.264	0.070	1.315	2.349	6.944	0.000	1.48
			0.310	0.041	0.002	0.229	0.390	7.492	0.000	



Note: Q=343.24; I^2=87.76%; T=0.07

3.4.1.2 How much does the effect size vary across studies (Heterogeneity)?

The Q-value is 343.236 with 42 degrees of freedom and $p<0.001$. Thus, it is recognised that the true effect size is not identical in all studies. 87.76% (I^2) of the variance in this observed effect is due to variance in true effects rather than sampling error. The T^2 is 0.069 representing the variance of true effect sizes and T is 0.262. The prediction interval is -0.13 to 0.75 therefore it would be expected the true effect size for 95% of all students receiving the formative assessment interventions to fall within this range.

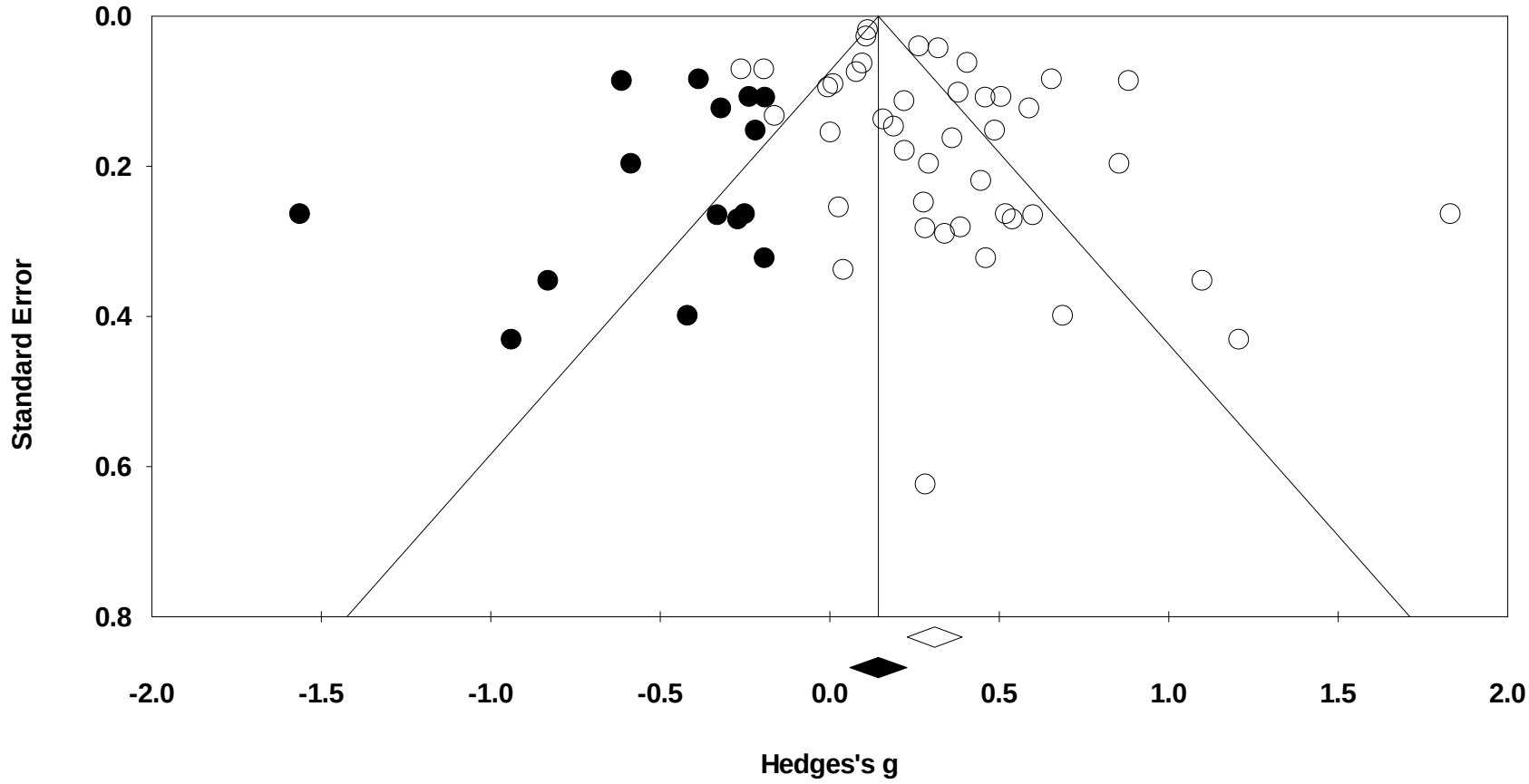
3.4.1.2. To what extent would publication bias or the small-study effect alter these findings?

The Classic fail-safe analysis showed that the incorporated data from the 43 observed effect sizes yielded a z-value of 16.41 and corresponding 2-tailed $p<0.0001$. The fail-safe N suggests that 2973 'null' effects would need to be included in the meta-analytic model for a combined 2-tailed of $p>0.05$ (i.e., for the reported effect to be nullified).

Duval and Tweedie's (2000) 'Trim and Fill' results suggest that 14 small studies should be trimmed from the left of the mean to determine the true 'centre' of the funnel plot and replace them with missing equivalents around the centre. The adjusted standardised effect size of this publication bias correction reduced the reported effect size from $g=0.31$ to $g=0.16$ (See Figure 3).

Figure 3

Funnel plot of standard error with observed and imputed studies by standardised difference in means (Hedges' g) (Combined Student Learning effects)



Note: White=Observed studies & effect; Black=Imputed studies & adjusted effect

3.4.2. Cognitive learning effects of formative assessment

There were 29 studies reporting formative assessment intervention strategies that sought to improve cognitive learning or development, therefore 31 effect sizes are captured in this meta-analysis. Of these 29 studies, 21% (*n*=6) of the papers met five or more of the quality assessment criteria.

3.4.2.1 Which formative assessment interventions most affect students' cognitive learning and development?

The included formative assessment interventions had a significant positive effect on cognitive outcomes, Z = 6.465, $p<0.001$. The standardised difference in means adjusting for sample size for all cognitive interventions was $g=0.29$ (95% CI [0.20, 0.38]; see Figure 4), meaning that on average, students receiving an intervention improved their cognitive learning or development by nearly a third of a standard deviation compared with those students who did not receive the same intervention.

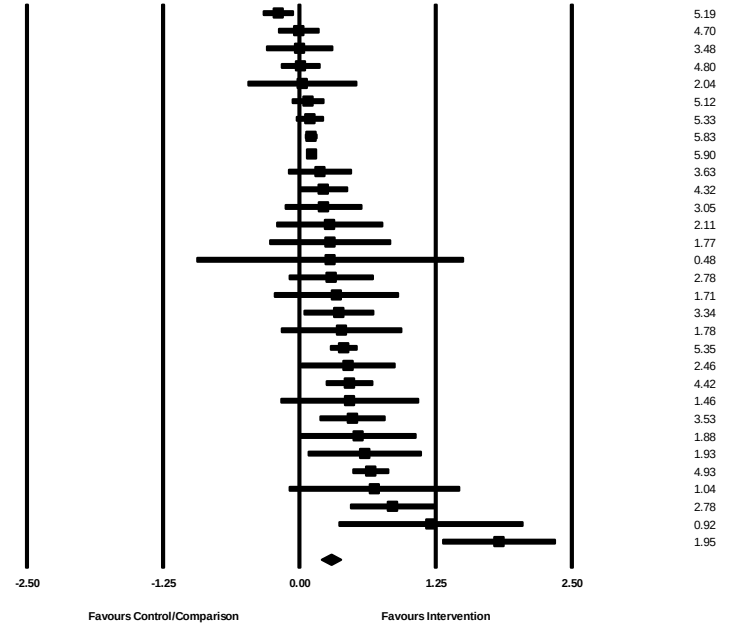
There were 16 studies that reported having a sample size adjusted effect size greater than the domain average of $g=0.29$. Five studies utilised computer-based feedback (Bush, 2021; Gijlers et al, 2013; Kegel & Bus, 2012; Koedinger et al, 2010; Vidal-Abarca et al, 2014) three incorporated self-assessment (Andrade et al, 2008; Gaintza & Goikoetxea, 2016; Gedikli & Buldur, 2022), two used peer assessment (Alitto et al, 2016; Gaintza & Goikoetxea, 2016) or performance feedback (Heir & Eckert, 2014; Truckenmiller et al, 2014), and the remainder used graphic organisers (Tavsanli et al, 2021), individualised

feedback (Damhuis et al, 2016), model-corrective feedback (Luquin & Garcia Mayo, 2021), or curriculum embedded formative assessment practices (Yin et al, 2014).

Figure 4

Cognitive effect of formative assessment by standardised difference in means (Hedges' g)

Study name	Subgroup within study	Outcome	Hedges's g	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value	Relative weight
Forster et al 2018	Combined	Reading	-0.193	0.071	0.005	-0.332	-0.054	-2.724	0.006	5.19
Decristan et al 2015b	Peer Assessment	Science Item Difficulty	-0.005	0.095	0.009	-0.191	0.181	-0.053	0.958	4.70
Fidalgo et al 2015	Combined	Writing quality	0.002	0.155	0.024	-0.302	0.306	0.014	0.989	3.48
Decristan et al 2015a	Formative Assessment	Science Item Difficulty	0.011	0.091	0.008	-0.167	0.188	0.119	0.905	4.80
Sukhram et al 2017	Fluency	Reading	0.027	0.255	0.065	-0.473	0.527	0.106	0.916	2.04
Koedinger et al 2010a	Regular	Mathematic achievement	0.080	0.075	0.006	-0.066	0.226	1.068	0.286	5.12
Gamlem et al 2019a	Norwegian National Test	Mathematic achievement	0.097	0.063	0.004	-0.027	0.221	1.536	0.124	5.33
Gamlem et al 2019b	Combined	Mathematic affective learning	0.108	0.027	0.001	0.055	0.161	3.985	0.000	5.83
Meyen et al 2010	Combined	Mathematic achievement	0.113	0.019	0.000	0.076	0.149	6.045	0.000	5.90
Prinsen et al 2013	Combined	Reading achievement	0.189	0.147	0.022	-0.099	0.478	1.287	0.198	3.63
Simmons et al 2015	Combined	Literacy	0.221	0.113	0.013	-0.001	0.442	1.950	0.051	4.32
Damhuis et al 2015	Combined	Vocabulary Acquisition	0.221	0.179	0.032	-0.130	0.573	1.233	0.217	3.05
Caceres et al 2021	Mathematic test	Mathematic achievement	0.278	0.248	0.062	-0.209	0.764	1.118	0.263	2.11
Fyfe et al 2016	Immediate feedback	Mathematic problem solving	0.282	0.283	0.080	-0.272	0.836	0.998	0.318	1.77
Bottge et al 2021	Mathematics test	Mathematic achievement	0.283	0.624	0.389	-0.940	1.505	0.453	0.650	0.48
Hier et al 2014	CBM-WE	Written expression	0.293	0.197	0.039	-0.093	0.678	1.487	0.137	2.78
Truckenmiller et al 2014	Correct writing sequence	Writing	0.340	0.290	0.084	-0.228	0.908	1.172	0.241	1.71
Kegel et al 2012	Tudor feedback	Literacy skills	0.362	0.163	0.026	0.043	0.681	2.223	0.026	3.34
Yin et al 2014	Exam performance	Science achievement	0.387	0.281	0.079	-0.165	0.938	1.374	0.169	1.78
Bush 2021	Mathematics test	Mathematic achievement	0.406	0.062	0.004	0.284	0.528	6.528	0.000	5.35
Gaintza et al 2016	Combined	Spelling	0.447	0.219	0.048	0.017	0.877	2.036	0.042	2.46
Alitto et al 2016	Combined	Writing achievement	0.460	0.109	0.012	0.247	0.672	4.234	0.000	4.42
Luqui et al 2021	Model feedback	Secondary Language Acquisition	0.461	0.322	0.104	-0.171	1.093	1.430	0.153	1.46
Koedinger et al 2010b	Special Ed	Mathematic achievement	0.488	0.152	0.023	0.189	0.786	3.200	0.001	3.53
Chen et al 2021	Reading test	Reading comprehension	0.540	0.271	0.073	0.009	1.070	1.992	0.046	1.88
Gijlers et al 2013	Combined	Science knowledge	0.600	0.265	0.070	0.080	1.120	2.263	0.024	1.93
Damhuis et al 2016	Combined	Vocabulary Acquisition	0.655	0.084	0.007	0.490	0.820	7.782	0.000	4.93
Vidal-Abarca et al 2014	PISA Test	Reading Literacy	0.688	0.399	0.159	-0.094	1.471	1.724	0.085	1.04
Andrade 2008	Essay scores	Writing achievement	0.855	0.197	0.039	0.469	1.240	4.345	0.000	2.78
Gedikli et al 2022	MAIfC	Scientific metacognition	1.208	0.431	0.186	0.363	2.052	2.802	0.005	0.92
Tavsanli et al 2021	Combined	Writing	1.832	0.264	0.070	1.315	2.349	6.944	0.000	1.95
			0.290	0.045	0.002	0.202	0.378	6.465	0.000	



Note: Q=177.09; I^2=83.06%; T=0.18

3.4.4.2 Heterogeneity of cognitive effects

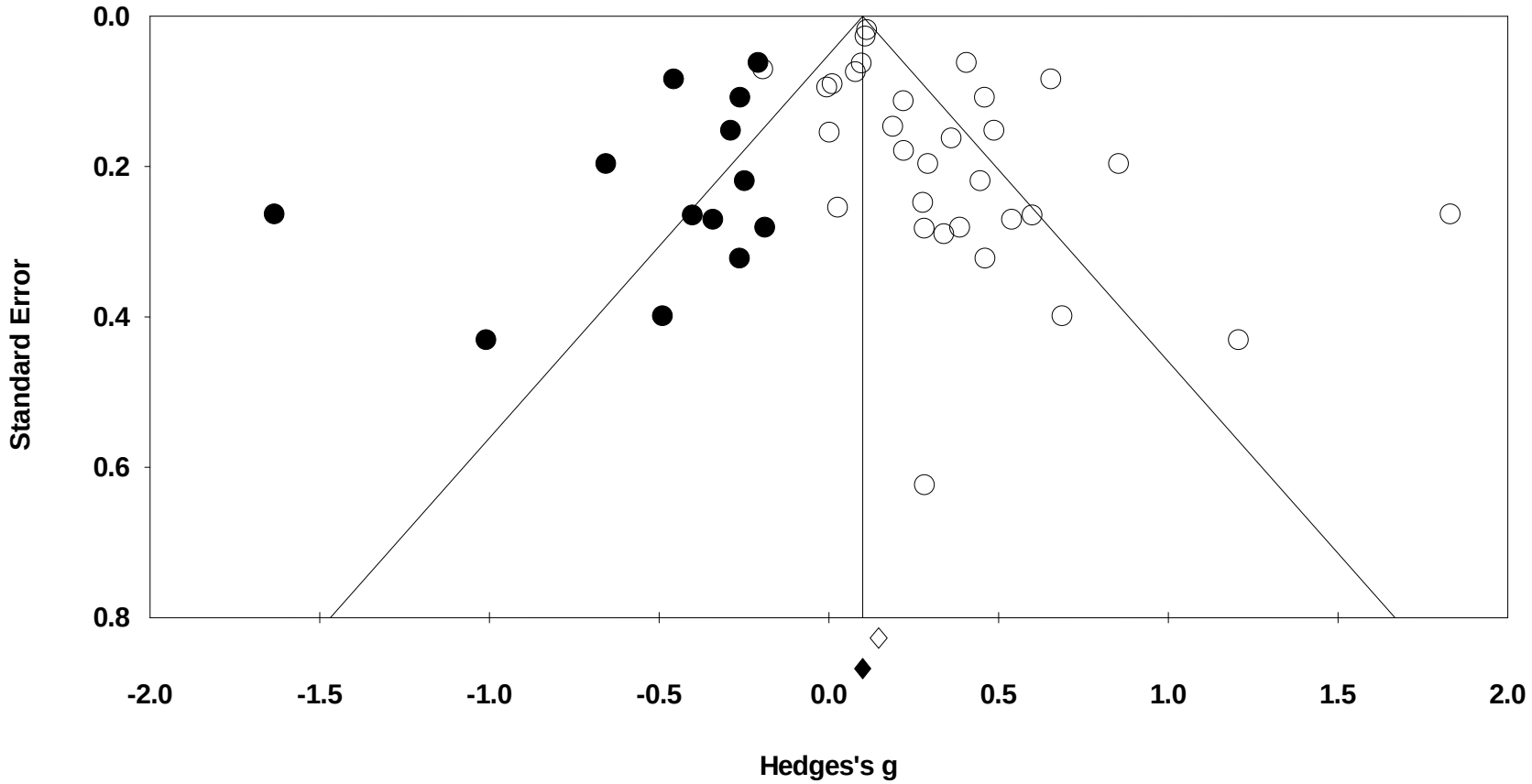
The Q-value for effects in the cognitive domain was 177.094 with 30 degrees of freedom and $p<0.001$. The I^2 is 83.06%, T^2 is 0.034 and T is 0.184. The prediction interval is -0.099 to 0.679 and therefore it is expected the true effect size for 95% of all students receiving a formative assessment intervention to improve cognitive learning and development to fall within this range.

3.4.4.3 To what extent would publication bias alter these findings?

The results of the Classic fail-safe analysis showed that the incorporated data from 31 effects yielded a z-value of 12.334 and corresponding 2-tailed $p<0.0001$. The fail-safe N in this case is 1197 suggesting that this many 'null' effects would need to be included for a combined 2-tailed $p>0.05$ (i.e., for the effect to be nullified). According to the 'Trim and Fill' analysis, 13 effects could be trimmed from the left of the mean to reduce the potential publication bias. The adjusted standardised effect size for formative assessment interventions on cognitive learning in this case would be decreased to $g=0.10$ or one tenth of a standard deviation (See Figure 5).

Figure 5

Funnel plot of standard error with observed and imputed studies by standardised difference in means (Hedges' g) (Cognitive Student Learning effects)



Note: White=Observed studies & effect; Black=Imputed studies & adjusted effect

3.4.4. Psychomotor learning effects of formative assessment

There were five (5) studies with five independent or combined effect sizes that examined the psychomotor learning and development occurring as the result of a formative assessment intervention. 40% (*n*=2) of these five papers met five or more of the quality assessment criteria.

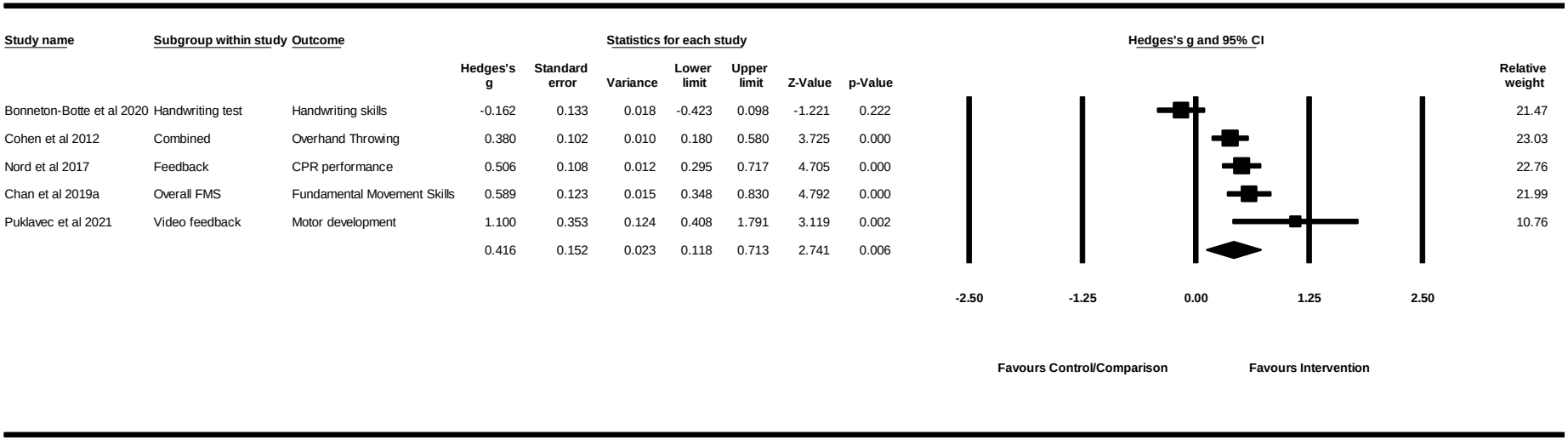
3.4.4.1 Which formative assessment interventions most affect student psychomotor learning and development?

The included formative assessment interventions had a significant positive effect on psychomotor outcomes, Z = 2.741, $p<0.001$. The standardised difference in means adjusting for sample size was $g=0.42$ (95% CI [0.118, 0.713]), meaning that students receiving the intervention improved their psychomotor learning and development by nearly a half of a standard deviation on average relative to those students who did not receive the intervention (See Figure 6).

Only three studies, each reporting a different formative assessment strategy intervention, reported a sample adjusted effect size greater than the psychomotor learning domain average of $g=0.42$. One study incorporated video feedback (Puklavec et al., 2021), one used curriculum-embedded assessment for learning practice (Chan et al., 2019), and one implemented immediate feedback (Nord et al., 2017).

Figure 6

Psychomotor effect of formative assessment by standardised difference in means (Hedges' g)



Note: Q=25.09; I^2=84.06%; T=0.30

3.4.4.2 Heterogeneity of psychomotor effects

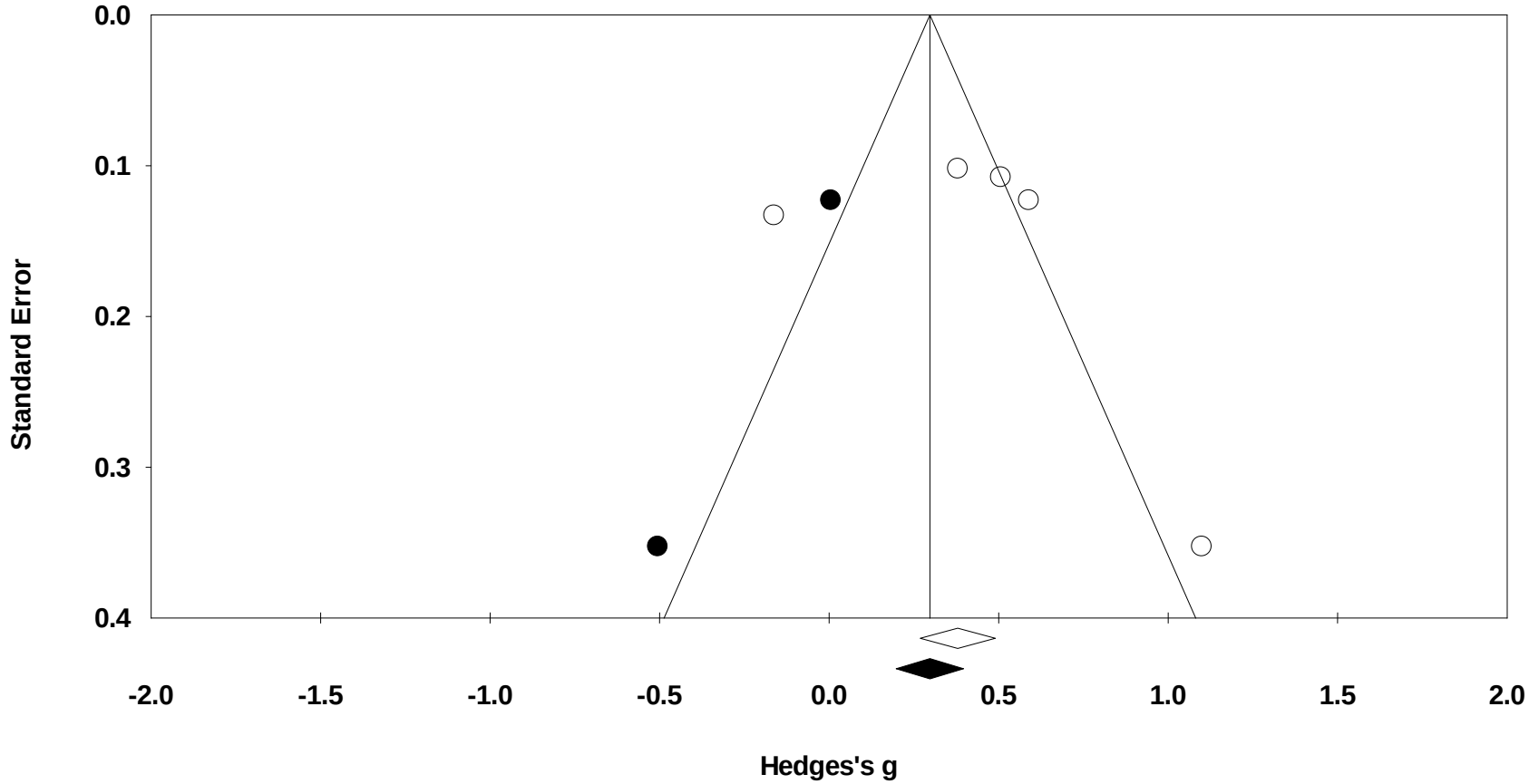
The Q-value for these effects was 25.089 with 4 degrees of freedom and $p<0.001$ ($I^2=84.06\%$; $T^2=0.089$; $T=0.299$). The prediction interval is -0.6489 to 1.4809 with the true effect size for 95% of all students receiving the interventions to fall within this range.

3.4.4.3 To what extent would publication bias alter these findings?

The results of the Classic fail-safe analysis showed that the incorporated data from effects yielded a z-value of 6.762 and corresponding 2-tailed $p<0.0001$. The fail-safe N in this case is 55. The 'Trim and Fill' analysis indicates two studies should be trimmed from the left of the mean to reduce the potential publication bias. This would result in an adjusted effect of $g=0.25$ (See Figure 7).

Figure 7

Funnel plot of standard error with observed and imputed studies by standardised difference in means (Hedges' g) (Psychomotor Student Learning effects)



Note: White=Observed studies & effect; Black=Imputed studies & adjusted effect

3.4.5. Affective learning effects of formative assessment

There were six (6) studies with seven (7) calculated effect sizes that examined the affective learning resulting from formative assessment interventions. 17% (*n*=1) of these papers met five or more of the methodological quality assessment criteria.

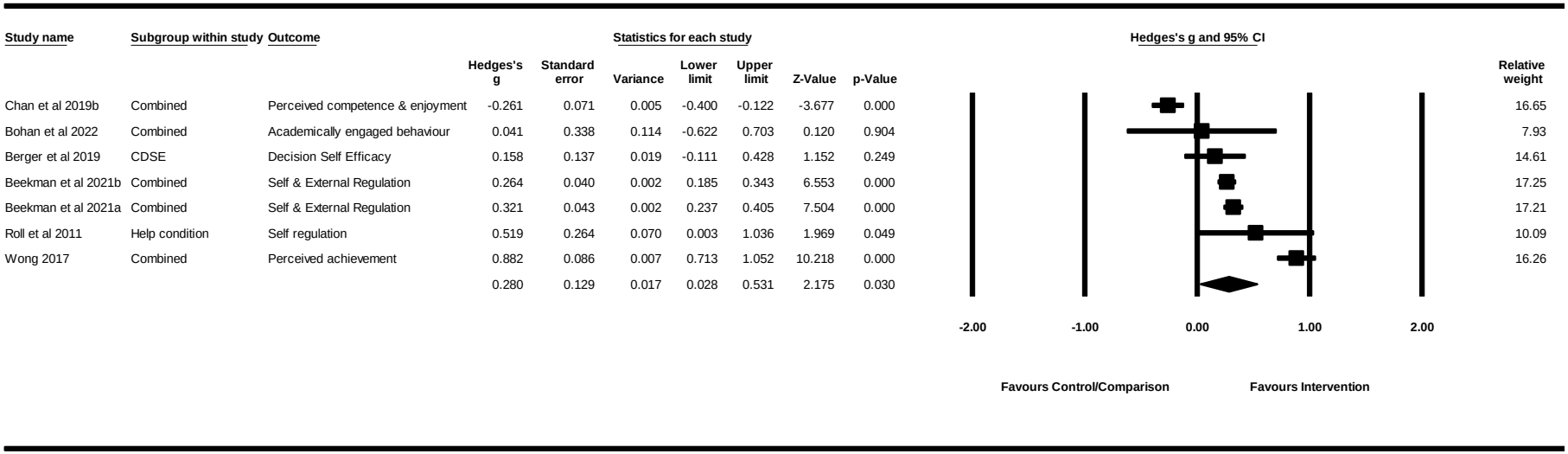
3.4.4.1 Which formative assessment interventions most affect student affective learning?

The included formative assessment interventions had a significant positive effect on affective learning, Z = 2.175, *p*=0.03. The standardised difference in means for all affective learning interventions when adjusting for sample size was *g*=0.28 (95% CI [0.028, 0.531]; see Figure 9), meaning that on average, students receiving the formative assessment intervention improved their affective learning by just over a quarter of a half of a standard deviation compared with those students who did not receive the intervention.

Again, only three studies, each reporting a different formative assessment strategy intervention, reported a sample adjusted effect size greater than the affective learning domain average of *g*=0.28. One study incorporated self-assessments (Wong, 2017), one used metacognitive feedback (Roll et al., 2011), and one implemented peer assessment (Beekman et al., 2021).

Figure 8

Affective effect of formative assessment by standardised difference in means (Hedges' g)



Note: Q=109.76; I^2=94.53%; T=0.31

3.4.4.2 Heterogeneity of affective effects

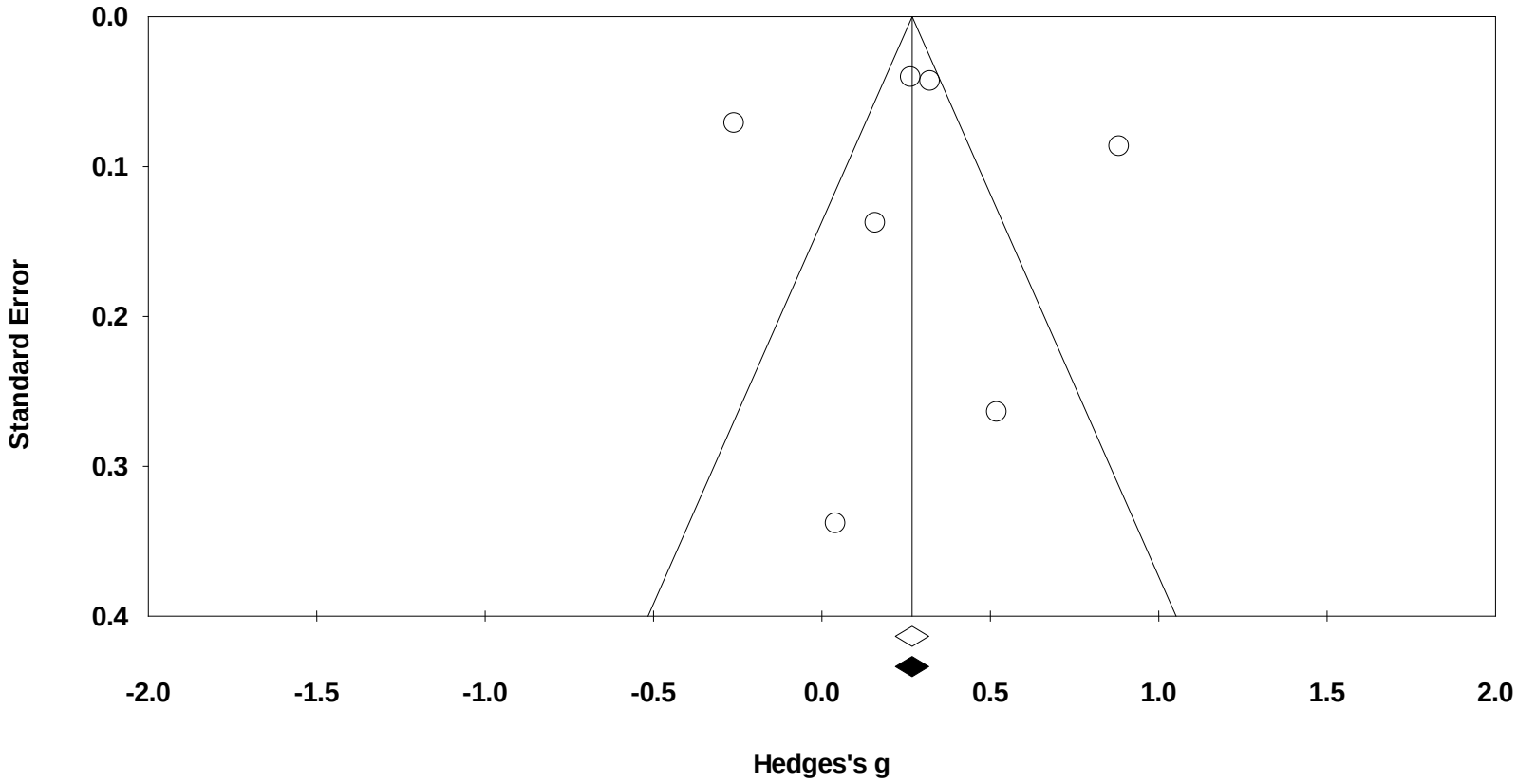
The Q-value for these effects is 109.762 with 6 degrees of freedom and $p<0.001$ ($I^2=94.534\%$; $T^2=0.094$; $T=0.307$). The prediction interval is -0.574 to 1.134 with the true effect size for 95% of all students receiving the interventions to fall within this range.

3.4.4.3 To what extent would publication bias alter these findings?

The results of the Classic fail-safe analysis showed that the incorporated data from 7 effects yielded a z-value of 9.01 and corresponding 2-tailed $p<0.0001$. The fail-safe N in this case is 141. The 'Trim and Fill' analysis indicates that no studies need to be trimmed from either side of the mean in order to achieve funnel plot symmetry, hence, no adjustment of the reported effect can be made for publication bias. (See Figure 9).

Figure 9

Funnel plot of standard error with observed and imputed studies by standardised difference in means (Hedges' g) (Affective Student Learning effects)



Note: White=Observed studies & effect; Black=Imputed studies & adjusted effect

CHAPTER 4

Discussion

The aim of the systematic review and meta-analysis was to systematically evaluate the impact of formative assessment on students' learning and development. Research question one sought to determine the sample size adjusted effect size, heterogeneity, and publication bias evident of formative assessment interventions on learning and development in primary and secondary school settings. The 39 studies found formative assessment had a significant effect on all learning and development domains of interest. Thus, not only can formative assessment operate as a mechanism for enhancing learning and development broadly, but specific interventions also have additive impacts over and above regular classroom instruction and pedagogy.

These findings are consistent with the earlier meta-analytic work of Kingston and Nash (2011), Graham et al (2015), Klute et al (2017), and Lee et al (2020) supporting the positive effect formative assessment has on student learning and development but indicates that these earlier reported effects may be inflated because they include studies of weaker design, do not adjust their effect sizes for sample size, fail to report on the multi-dimensional nature of learning and development, and do not report the inflationary effects of publication bias.

This systematic review has highlighted that the possible inflationary effect caused by the lack of methodological quality of formative assessment interventions warrants attention. Eight of the 10 methodological assessment criteria assessed in this systematic

review revealed that less than 50% of the studies meet the stated criterion. Five methodological assessment criteria reviewed in the pool of studies had less than a quarter of the reviewed studies achieve them. This highlights a significant concern in drawing inference into the effect that formative assessment has on student learning and development. Rutkowski and Delandshere (2016) identify interpretations made with a high degree of risk of bias and methodological shortcomings of educational assessment studies as the most significant constraint on drawing causal inference from educational interventions. They conclude that RCT, CT and QE designs of educational assessment interventions play a vital role in developing an understanding of essential phenomena in educational research. Furthermore, they contend that many such studies contribute to the increased understanding of educational systems and allow both policy makers and researchers to explore the effect of assessment in greater depth (Rutkowski & Delandshere, 2016). Whilst the use of RCTs, CTs, and QEs in educational assessment research remains a paradigmatic contention (Pampaka et al., 2016), this study does demonstrate that greater attention to the methodological quality of such approaches to educational research require far more attention if they are to continue influencing the decisions made around formative assessment as a viable education intervention strategy.

Speaking to the inflationary effects reported by formative assessment interventions, by only incorporating studies that were RCTs, CTs, and QEs in this meta-analysis along with examining different learning domains and adjusting for sample size, the overall standardised effect size of formative assessment interventions was g=0.31. This represents an effect size estimate at the lower range of the four meta-analytic studies previously published [Kingston & Nash, 2011 (d=0.32); Graham et al., 2015 (d=0.38); Klute et al.,

2017 (*d*=0.29); Lee et al., 2020 (*d*=0.29)] and below the lower margin of ES=0.4 claimed by the seminal Black and Wiliam (1997) paper.

However, when the accounting for publication bias, the adjusted effect size for formative assessment interventions on learning and development would be reduced to as little as *g*=0.16. This should serve as significant alarm for educators and researchers in this discipline about the need for a greater range of formative assessment studies to be reported in the empirical literature. As highlighted by Bennett (2011), it may be that the field has not been able to clearly articulate and agree on the scope of formative assessment strategies being used in education or that only specific high yield formative assessment strategies in certain educational disciplines are finding their way into empirical publications.

Chambers (2017) has written quite extensively on the causes of publication bias and the pursuit of affirmative and high yield findings in the education and psychology literature. He contends that the disposition of journals seeking to publish affirmative outcomes proliferates publication bias in the educational and psychological literature due to the inherent characteristics of conventional statistical analyses these researchers typically employ. Especially in RCTs, CT, and QE research designs. Specifically, Chambers (2017) argues that publication bias occurs because affirmative findings are actively seeking to reject the null hypothesis in support of the alternative hypothesis. Following this line of argument, applying null hypothesis significance testing as the statistical default framework only quantifies the probability of observing an effect of equal or greater magnitude under the assumption that the null hypothesis holds true. This approach however, does not estimate the probability of the null hypothesis itself being accurate. Instead, p-values can

only ever provide an estimation of the likelihood of observing the specific effect determined to be sufficient to qualify as a high yield outcome (based on the study's power or sample size), rather than determining the likelihood of a particular hypothesis being valid considering the observed effect.

Consequently, whilst a statistically significant finding permits an educational researcher to reject their null hypothesis, a statistically non-significant result does not provide grounds for the researcher to embrace it. In the case of statistically non-significant results, the researcher's conclusion is restricted to the possibility that null hypothesis may hold true, or that the data might lack the sensitivity to detect the observed effect determined to be sufficient to qualify as a high yield outcome. Therefore, drawing definitive conclusions from statistically non-significant findings inherently remains inconclusive and often not worthy of publication in leading education and psychology journals (Chambers, 2017).

Whilst the preference of journals towards disseminating affirmative findings constitutes a catalyst for the emergence of publication bias in the educational literature; nevertheless, it reveals merely part of the overarching issue. What else remains is the relentless pursuit of originality and/or novelty in educational interventions (Kristjánsson, 2012). To contend for coveted publication in high impact journals, submissions are compelled to either embrace an innovative methodological framework or yield a ground-breaking discovery, if not, both. Most of the empirical outlets that serve as conduits for educational research evaluate the worthiness of submitted manuscripts, at least partially, through the prism of novelty and originality (Chambers, 2017). Even a rapid review of

literature published in the last 12 months reveals that authors frequently affirm their use of a ‘novel-approach’ being undertaken in their research pertaining to formative assessment (i.e., Al-Zohbi et al., 2023; Licchelli & Barnett, 2023; Ochsen et al., 2023; Wilkie, Ayalon, & Kanj, 2023).

Another explanation for the decline in observed effect size attributed to formative assessment was accounted for in a study by Cheung and Slavin (2016). They argue that the growing significance of evidence in educational policy compels researchers to comprehend how research designs influence the reported effect sizes in experiments assessing educational programs. Their study delved into 645 research papers from 12 evaluation reviews of preschool, reading, mathematics, and science program interventions. Their findings report that all other study designs in this universe of educational literature exhibit significantly higher effect sizes compared to randomised and controlled study designs. Based on these findings, one of the five key recommendations the authors make to education policy makers and teachers is that they should assert that larger, randomised evaluations are being undertaken before system-wide policy or pedagogical decisions are made (Cheung & Slavin, 2016). Furthermore, Cheung and Slavin (2016) claim that weaker designed studies are seriously overstating the impact they are having on student learning. At a minimum, it may be that weaker designed interventions are allowing for a greater variation in outcomes to occur in the previous meta-analyses of formative assessment. If critical decisions to implement formative assessment into educational policy and practice are being pursued because they are based on evidence, that evidence should be as conclusive, and with as little variation, as possible.

To understand this phenomenon of diminishing effect sizes further, research questions two and three examined the extent to which the reported effect size, heterogeneity, and publication bias of formative assessment on learning and development were constrained by learning domain. They also examined the reported effect size estimates of specific formative assessment practices that were being conducted in both primary and secondary school settings.

Case in point, this meta-analysis found that of all the studies that reported effect sizes greater than g=0.31, seven studies focussed exclusively on spelling, writing achievement or written expression (Alitto et al 2016; Andrade 2008; Fidalgo et al 2015; Gaintza et al 2016; Heir et al 2014; Tavsanli et al 2021; Truckenmiller et al 2014). On average, these studies reported a sample size adjusted effect of g=0.58. Furthermore, when these studies were analysed for publication bias using the Trim and Fill Method (Duval & Tweedie, 2000), two studies were required to be trimmed from the right of the mean before plot symmetry was achieved and therefore increased the reported effect size estimate to g=0.77. Both effect size estimates concur with the range initially expressed by Black and Wiliam's (1997) seminal work.

Conversely, of the five studies that focussed exclusively on reading-related learning outcomes (Chen et al 2021; Forster et al 2018; Prinsen et al 2013; Sukhram et al 2017; Vidal-Abarca et al 2014), these studies reported a sample size adjusted effect of just g=0.16. More alarming is that when these studies were analysed for publication bias, the adjusted effect size estimate drops to effectively zero (g=-0.01).

The importance of effect sizes for comparing results across different measures, groups, and intervention designs is a significant component of conducting an informative meta-analysis (Hattie, Rogers, & Swaminathan, 2014). To further explore the impact of formative assessment interventions on students' learning and development more comprehensively, this study conducted an analysis to examine the relative effect sizes both between and within learning domains, specifically the psychomotor, cognitive, and affective learning domains. A recent meta-analysis by Dudley and colleagues (2022) examined the effect of physical education-based interventions on psychomotor learning and development of students. They concluded that the average standardised effect of physical education interventions on psychomotor development to be d=0.52 based on 51 studies published over 25 years. Whilst this current systematic review and meta-analysis only uncovered five formative assessment interventions to impact on psychomotor learning, the sample size adjusted effect size of g=0.42 suggests that formative assessment practices have the potential to make a worthwhile contribution to a student's psychomotor learning. However, the inflationary effects of publication bias and methodological quality of these formative assessment approaches in developing psychomotor learning should also be addressed.

Whilst it is difficult to ascertain the comparative effect of formative assessment interventions on the affective learning of students due to a dearth of literature in the field, it is possible to conduct comparisons with effects in the cognitive learning domain. In Hattie's (2009) seminal analysis of the school, parent, student, and teacher impacts to student academic achievement outcomes, the average effect size across all meta-analyses conducted equated to d=0.4 (Hattie, 2009). Whilst Hattie's analysis did not investigate

formative assessment interventions specifically, he did report on some of the strategies and student groups that were captured in this systematic review as being implemented as formative assessment interventions.

Case in point, Hattie (2010) cites Fuchs and Fuchs (1986) study as a seminal work that substantiates systematic formative evaluation being conducted by teachers. The reported standardised effect of systematic formative evaluation by teachers working with children with a mild intellectual disability was d=0.70. By comparison, this meta-analysis (based on five reported effect sizes from four different studies) found that the effect size estimate of formative assessment interventions on students with disabilities was only g=0.12 (Q = 8.691; I^2 = 53.97%, p=0.07). Although a great deal of heterogeneity exists in this pool of studies, the disparity between the reported effect size estimates is sufficient to warrant scrutiny as to whether formative assessment strategies are a worthwhile educational investment for students who experience a disability.

One of John Hattie's strongest claims in Visible Learning (Hattie, 2009) is that feedback is one of the most powerful influences on student cognitive learning and development. He reports that the average standardised effect drawn from meta-analyses on student achievement as the result of feedback is d=0.73, or an improvement of nearly three quarters of a standard deviation compared to those students who don't receive feedback (Hattie, 2009; p.173). This statistic is drawn from meta-analyses capturing some 1,287 studies. Whilst Hattie (2009) acknowledges the considerable variability in the approaches to feedback, this meta-analysis, based on 22 studies, reported feedback-based formative assessments resulted in a sample size adjusted effect size of just g=0.33 (Q=97.199;

I^2=78.40%, p<0.01). Given the disparity in these findings, greater scrutiny should be paid in future intervention into the sources and associated pedagogy in which feedback is deployed through formative assessments.

An additional layer of complexity in this meta-analysis, missing from other meta-analyses of formative assessment interventions to date, is that this data was of sufficient size and clarity to disaggregate by learning domain. Importantly, and notwithstanding the positive overall effects, this study found variation in the magnitude of specific interventions both between and within our domains of interest. It was found that only in the cognitive learning domain were there consistent approaches that demonstrated effectiveness above the average effect size estimate. However, in other learning domains (i.e., the psychomotor and affective learning domains), no formative assessment approaches were identified as being consistently greater than the combined average effect size estimate or of that within their respective learning domain. This can be accounted for due to the relative low number of formative assessment studies that have examined psychomotor and affective learning outcomes. However, in terms of the cognitive domain, one specific formative assessment strategy does consistently report above average effects.

There were five studies that utilised computer-based feedback approaches to formative assessment (Bush, 2021; Gijlers et at., 2013; Kegel & Bus, 2012; Koedinger et al., 2010; Vidal-Abarca et al., 2014) and they reported a sample size adjusted effect size greater than the cognitive domain average (i.e., g>0.29). Closer scrutiny of the five computer-based feedback formative assessment interventions that yielded the higher-than-average effect sizes reveal all five assessed students in numeracy (Bush, 2021; Koedinger et

al., 2010), literacy (Kegel & Bus, 2012; Vidal-Abarca et al., 2014), or science (Gijlers et at., 2013) learning outcomes. This finding is supported by an earlier meta-analysis published by Van der Kleij and colleagues (Van der Kleij et al., 2015) that investigated the effects of feedback in computer-based learning environments on student learning outcomes. The authors of this earlier meta-analysis conclude that comparable effect sizes were found for mathematics (d=0.93), science (d=0.40), and language (d=0.25) learning outcomes when computer-assisted feedback systems are employed. Furthermore, their meta-analysis revealed that larger effect sizes are reported by computer-based feedback systems when they provide elaborated feedback to students rather than simply knowledge of results or knowledge of correct response (Van der Kleij et al., 2015). Four of the five computer-based feedback interventions in this meta-analysis (Bush, 2021; Kegel & Bus, 2012; Koedinger et al., 2010; Vidal-Abarca et al., 2014) explicitly state that their formative assessment processes entailed elaborated feedback via personalised tutoring or a procedural feedback strategy.

This systematic review also revealed that peer assessment remains a widely recognised strategy in the field of formative assessment and this continues to be supported in the existing literature base (Yan & Boud, 2021). This type of formative assessment strategy has garnered significant attention primarily because it is claimed to be able to foster student-centred learning (Tai and Sevenhuysen, 2018; Xiang et al., 2022).

While it is generally believed that peer assessment positively impacts learning outcomes in primary and secondary school students, the findings from a recent meta-analysis on the topic yielded mixed results. Li and colleagues (2019) conducted a meta-

analysis that synthesised data from 58 studies, encompassing 134 effect sizes on the effects to which peer-assessment promoted student learning. Their findings reported that students who engaged in peer assessment demonstrated a notable increase in their performance, with a d=0.29 improvement compared to those who do not participate in such activities.

Interestingly, the four formative assessment interventions that utilised a peer-assessment strategy in this study (Alitto et al., 2016; Beekman et al., 2021; Fidalgo et al., 2015; Prinsen et al., 2013) yielded a comparable effect size of g=0.28, even when adjusted for sample size. Li et al (2019) argues that rater training is the most influential factor in peer-assessment efficacy. They claim that students who receive training in assessing their peers will exhibit significantly larger effects in peer assessment compared to those who do not receive such training. Three of four formative assessment interventions in this study included a period of rater training for students undertaking peer assessments (Alitto et al., 2016; Fidalgo et al., 2015; Prinsen et al., 2013) yet their reported effect sizes vary greatly between a medium effect of g=0.46 (Alitto et al., 2016) down to a negligible g=0.002 (Fidalgo et al., 2015)

Moreover, Li et al (2019) reported that computer-mediated peer assessment exhibited greater learning gains when compared to paper-based peer assessment. This study challenges this contention regarding formative assessment that utilises a peer assessment strategy. Whilst the number of studies are limited, only one study was uncovered in this review that employed a computer-assisted peer assessment procedure (Prinsen et al., 2013) and its reported effect on student learning was only g=0.19. When compared to the two

studies that utilised paper-based strategies (Alitto et al., 2016; Beekman et al., 2021), they reported sample size adjusted learning effects of g=0.47 and g=0.32 respectively.

Limitations

Although the current meta-analysis provided valuable insights into the impact of formative assessment interventions on the development of our targeted learning domains, there are noteworthy limitations to consider. Firstly, this study did not explore the dosage or duration of interventions for any of the specific intervention methods. Consequently, when comparing the different pedagogical approaches to formative assessment, it is essential to recognise that the results may be influenced not only by the specific approach itself but also by the implementation of the program. The evidence regarding the significance of dosage in this context is inconclusive. Future research should aim to disentangle the effects of individual formative assessment strategies and their frequency to ascertain their independent contributions as well as their combined impact on student learning. By conducting studies that systematically vary these factors, we can gain a deeper understanding of the unique benefits offered by different strategies and determine the optimal frequency at which they should be implemented to maximize their effectiveness. Such investigations would provide valuable insights for educators and policymakers seeking evidence-based approaches to enhance student learning outcomes.

Furthermore, in studies involving multiple longitudinal follow-up assessments, it was crucial to select a single measurement point after the intervention. In such cases, the last assessment completed was chosen as the reference point. To gain a comprehensive

understanding of the efficacy and durability of promising formative assessment interventions, it is advisable for future research to examine the trajectory of change over time within each of the three learning domains. This will provide insights into the extent and duration of the intervention's effect.

Additionally, it is important to note that the evidence presented in this analysis was not weighted based on a scientific hierarchy, such as prioritising RCTs, CTs, or QEs. However, the assessment of methodological quality provides an opportunity to evaluate the robustness of the reported study details and should be considered when interpreting the results. To gain further insights and enhance understanding of potential sampling biases, future research should consider analysing and comparing these different study designs separately as the body of literature grows, allowing for distinct conclusions to be drawn for each design.

Also, it is important to acknowledge that the instruments engaged to gather data for each of the three domains varied among the studies included in this meta-analysis. Some studies utilised subjective instruments, while others relied on objective measures to assess identical constructs. Moreover, the constructs themselves exhibited variations within each domain. Additionally, as previously discussed, certain domains inherently present greater challenges and relative efficacy compared to others.

Lastly, it is worth noting that specific outcome measures within each domain may differ in their sensitivity to change. Consequently, when interpreting the broad statistical

trends reported in this analysis, it is essential to exercise caution and consider the conceptual and experimental differences that exist across studies.

Conclusion

In conclusion, this dissertation has provided comprehensive insights into the effects of formative assessment interventions and strategies aimed at optimising learning outcomes in children and adolescents. Through a systematic review and meta-analysis, 38 studies encompassing over 16,000 participants were analysed, resulting in 43 calculated effect sizes.

The findings indicate that formative assessment interventions have a small to medium-sized positive effect on learning outcomes, with a sample size adjusted mean effect of g=0.31 (95% CI: 0.29 to 0.39). However, when adjusted for publication bias using the Trim and Fill Method, the mean effect size decreases to as little as g=0.16. Furthermore, it is important to note that the reported effects vary significantly depending on the learning outcomes and the learning domains being targeted. Additionally, less than 10% of the formative assessment interventions included in this study reported a negative effect on student learning.

Overall, this thesis concludes that formative assessment interventions consistently yield positive effects on student learning in both primary and secondary school settings across all three learning domains. However, it highlights the potential inflation of reported effects in previous meta-analytic studies due to the inclusion of weaker study designs.

Moreover, the thesis emphasises the limitations of the evidence reported, including poor methodological quality and substantial publication bias in the included studies.

It is recommended that future research addresses these limitations by employing stronger study designs and addressing publication bias issues. Furthermore, studies should consider the consistency in intervention dosage, study design, and data collection instruments to facilitate more accurate and comparable findings. By doing so, researchers and educators can further enhance the implementation and effectiveness of formative assessment strategies to optimise student learning and development.

References

Al-Zohbi, G., Pilotti, M. A., Barghout, K., Elmoussa, O., & Abdelsalam, H. (2023). Lesson learned from the pandemic for learning physics. *Journal of Computer Assisted Learning, 39*(2), 591-602.

Babineau, J. (2014). Product review: Covidence (systematic review software). *Journal of the Canadian Health Libraries Association/Journal de l'Association des bibliothèques de la santé du Canada, 35*(2), 68-71.

Baxter Magolda, M. B. (2000). Teaching to promote holistic learning and development. *New directions for teaching and learning, 2000*(82), 88-98.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in education: principles, policy & practice, 18*(1), 5-25.

Bhagat, K. K., & Spector, J. M. (2017). Formative assessment in complex problem-solving domains: The emerging role of assessment technologies. *Journal of Educational Technology & Society, 20*(4), 312-317.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-148.

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

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis.* John Wiley & Sons.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice, 30*(1), 3-12.

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.

Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher, 45*(5), 283–292.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10,* 101-129.

Cotton, W., Dudley, D., Peralta, L., & Werkhoven, T. (2020). The effect of teacher-delivered nutrition education programs on elementary-aged students: An updated systematic review and meta-analysis. *Preventive Medicine Reports, 20,* 101178.

Darling-Hammond, L., & McLaughlin, M. W. (2011). Policies that support professional development in an era of reform. *Phi Delta Kappan, 92*(6), 81-92.

DerSimonian, R., & Laird, N. (2015). Meta-analysis in clinical trials revisited. *Contemporary Clinical Trials*, *45*, 139-145.

Dorn, S. (2010). The political dilemmas of formative assessment. *Exceptional Children, 76*(3), 325–337

Dudley, D., & Burden, R. (2020). What effect on learning does increasing the proportion of curriculum time allocated to physical education have? A systematic review and meta-analysis. *European Physical Education Review*, *26*(1), 85-100.

Dudley, D. A., Cotton, W. G., & Peralta, L. R. (2015). Teaching approaches and strategies that promote healthy eating in primary school children: a systematic review and meta-analysis. *International Journal of Behavioral Nutrition and Physical Activity*, *12*, 1-26.

Dudley, D., Mackenzie, E., Van Bergen, P., Cairney, J., & Barnett, L. (2022). What drives quality physical education? A systematic review and meta-analysis of learning and development effects from physical education-based interventions. *Frontiers in Psychology*, 3177.

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessments: The limited scientific evidence of the impact of formative assessments in education. *Practical Assessment, Research, and Evaluation, 14*(1), 7.

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.

Dweck, C. S. (2006). *Mindset: The new psychology of success*. Random House.

Elmahdi, I., Al-Hattami, A., & Fawzi, H. (2018). Using Technology for Formative Assessment to Improve Students' Learning. *Turkish Online Journal of Educational Technology-TOJET*, *17*(2), 182-188.

Forrest, S. (2018). Can CPD enhance student-centred teaching and encourage explicit instruction of International Baccalaureate approaches to learning skills? A qualitative formative assessment and summative evaluation of an IB School's In-House CPD Programme. *Journal of Research in International Education*, *17*(3), 262-285.

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional children*, *53*(3), 199-208.

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

Guskey, T. R. (2007). Closing the knowledge gap on effective professional development. *Educational Horizons*, *85*(2), 74-79.

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

Hattie, J., Rogers, H. J., & Swaminathan, H. (2014). The role of meta-analysis in educational research. *A Companion to Research in Education*, 197-207.

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

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107-128.

Heritage, M. (2007). Formative assessment: What do teachers need to know and do?. *Phi Delta Kappan, 89*(2), 140-145.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*, 1539-1558.

Hinde, S., & Spackman, E. (2015). Bidirectional citation searching to completion: an exploration of literature searching methods. *Pharmacoeconomics, 33*, 5-11.

Hoque, M. E. (2016). Three domains of learning: Cognitive, affective and psychomotor. *The Journal of EFL Education and Research, 2*(2), 45-52.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 28-37.

Klute, M., Apthorp, H., Harlacher, J., & Reale, M. (2017). *Formative Assessment and Elementary School Student Academic Achievement: A Review of the Evidence*. Regional Educational Laboratory Central.

Kolb, A. Y., & Kolb, D. A. (2009). Experiential learning theory: A dynamic, holistic approach to management learning, education and development. *The SAGE handbook of management learning, education and development*, *7*, 42.

Kristjánsson, K. (2012). Positive psychology and positive education: Old wine in new bottles?. *Educational psychologist*, *47*(2), 86-105.

Martinez, J. G. R., & Martinez, N. C. (1992). Re-examining repeated testing and teacher effects in a remedial mathematics course. *British Journal of Educational Psychology, 62*, 356-363.

Li, H., Xiong, Y., Hunter, C. V., Guo, X., & Tywoniw, R. (2020). Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education, 45*(2), 193-211.

Licchelli, S., & Barnett, L. (2023). Using an online escape room as a formative assessment tool during a lecture on HIV: a case study. *Journal of Learning Development in Higher Education*, (27), 1-14.

Lin L., & Aloe A. M. (2021). Evaluation of various estimators for standardized mean difference in meta-analysis. *Statistics in Medicine, 40*, 403–426.

Lin, L., & Chu, H. (2018). Quantifying publication bias in meta-analysis. *Biometrics, 74*(3), 785-794.

Moss, C. M., & Brookhart, S. M. (2019). *Advancing formative assessment in every classroom: A guide for instructional leaders*. ASCD.

Ochsen, S., Bernholt, A., Grund, S., & Bernholt, S. (2023). Interestingness is in the eye of the beholder–the impact of formative assessment on students' situational interest in chemistry classrooms. *International Journal of Science Education*, *45*(5), 383-404.

Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, *8*(2), 157-159.

Ozan, C., & Kincal, R. (2018). The effects of formative assessment on academic achievement, attitudes toward the lesson, and self-regulation skills. *Educational Sciences-Theory & Practice, 18*(1), 85-118.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement. *Journal of Clinical Epidemiology*, *134*, 103-112.

Pampaka, M., Williams, J., & Homer, M. (2016). Is the educational 'what works' agenda working? Critical methodological developments. *International Journal of Research & Method in Education, 39*(3), 231-236.

Parrila, R., Dudley, D., Song, S., & Georgiou, G. K. (2020). A meta-analysis of reading-level match dyslexia studies in consistent alphabetic orthographies. *Annals of Dyslexia*, *70*, 1-26.

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

Rücker, G., Schwarzer, G., Carpenter, J. R., & Schumacher, M. (2008). Undue reliance on I^2 in assessing heterogeneity may mislead. *BMC Medical Research Methodology*, *8*, 1-9.

Rutkowski, D., & Delandshere, G. (2016). Causal inferences with large scale assessment data: Using a validity framework. *Large-scale Assessments in Education, 4*(1), 1-18.

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

Shepard, L. A. (2010). Commentary: Evaluating the validity of formative and interim assessment. *Educational Measurement: Issues and Practice, 28*(3), 32–37.

Tai, J., & Sevenhuysen, S. (2018). The role of peers in developing evaluative judgement. *In Developing Evaluative Judgement in Higher Education (pp. 156-165)*. Routledge.

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of Feedback in a Computer-Based Learning Environment on Students' Learning Outcomes: A Meta-Analysis. *Review of Educational Research, 85*(4), 475–511.

Van Sluijs, E. M., McMinn, A. M., & Griffin, S. J. (2007). Effectiveness of interventions to promote physical activity in children and adolescents: systematic review of controlled trials. *British Medical Journal*, *335*(7622), 703.

West, D. M. (2012). Big data for education: Data mining, data analytics, and web dashboards. *Governance Studies at Brookings, 4*(1), 1-10

Wiliam, D. (2011). *Embedded formative assessment*. Solution Tree Press.

Wilkie, K. J., Ayalon, M., & Kanj, S. Z. (2023). Exploring ways to engage disaffected mathematics students through formative assessment processes with rich tasks. *Teaching and Teacher Education, 132*, 104256.

Wylie, C., & Lyon, C. (2016). *Using the formative assessment rubrics, reflection and observation tools to support professional reflection on practice (Revised).* Formative Assessment for Students and Teachers.

Xiang, X., Yuan, R., & Yu, B. (2022). Implementing assessment as learning in the L2 writing classroom: a Chinese case. *Assessment & Evaluation in Higher Education, 47*(5), 727-741.

Yan, Z., & Boud, D. (2021). Conceptualising assessment-as-learning. *In Assessment as Learning (pp. 11-24).* Routledge.

Yeh, S. (2010). Understanding and addressing the achievement gap through individualized instruction and formative assessment. *Assessment in Education: Principles, Policy & Practice, 17*,169–182.

References included in the systematic review and/or meta-analysis

Alitto, J., Malecki, C. K., Coyle, S., & Santuzzi, A. (2016). Examining the effects of adult and peer mediated goal setting and feedback interventions for writing: Two studies. *Journal of School Psychology*, *56*, 89-109.

Andrade, H. L., Du, Y., & Wang, X. (2008). Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing. *Educational Measurement: Issues and Practice*, *27*(2), 3-13.

Beekman, K., Brinke, J. T., & Boshuizen, E. (2021). Sustainability of developed self-regulation by means of formative assessment among young adolescents: A longitudinal study. In *Frontiers in Education* (p. 444). Frontiers.

Berger, N., Hanham, J., Stevens, C. J., & Holmes, K. (2019). Immediate feedback improves career decision self-efficacy and aspirational alignment. *Frontiers in Psychology*, *10*, 255.

Bohan, C., McDowell, C., & Smyth, S. (2022). Does the immediacy of feedback matter in game-based classroom management? Analysis of the caught being good game with adolescent students. *Journal of Positive Behavior Interventions*, *24*(3), 208-221.

Bonneton-Botté, N., Fleury, S., Girard, N., Le Magadou, M., Cherbonnier, A., Renault, M., ... & Jamet, E. (2020). Can tablet apps support the learning of handwriting? An

investigation of learning outcomes in kindergarten classroom. *Computers & Education*, *151*, 103831.

Bottge, B. A., Ma, X., Gassaway, L. J., Jones, M., & Gravil, M. (2021). Effects of formative assessment strategies on the fractions computation skills of students with disabilities. *Remedial and Special Education*, *42*(5), 279-289.

Bush, J. B. (2021). Software-based intervention with digital manipulatives to support student conceptual understandings of fractions. *British Journal of Educational Technology*, *52*(6), 2299-2318.

Cáceres, M., Nussbaum, M., González, F., & Gardulski, V. (2021). Is more detailed feedback better for problem-solving?. *Interactive Learning Environments*, *29*(7), 1189-1210.

Chan, C. H., Ha, A. S., Ng, J. Y., & Lubans, D. R. (2019). The A+ FMS cluster randomized controlled trial: An assessment-based intervention on fundamental movement skills and psychosocial outcomes in primary schoolchildren. *Journal of Science and Medicine in Sport*, *22*(8), 935-940.

Chen, C. M., Chen, L. C., & Horng, W. J. (2021). A collaborative reading annotation system with formative assessment and feedback mechanisms to promote digital reading performance. *Interactive Learning Environments*, *29*(5), 848-865.

Cohen, R., Goodway, J. D., & Lidor, R. (2012). The effectiveness of aligned developmental feedback on the overhand throw in third-grade students. *Physical Education and Sport Pedagogy*, *17*(5), 525-541.

Damhuis, C. M., Segers, E., & Verhoeven, L. (2015). Stimulating breadth and depth of vocabulary via repeated storybook readings or tests. *School Effectiveness and School Improvement*, *26*(3), 382-396.

Damhuis, C. M., Segers, E., Scheltinga, F., & Verhoeven, L. (2016). Effects of individualized word retrieval in kindergarten vocabulary intervention. *School Effectiveness and School Improvement*, *27*(3), 441-454.

Dao, P., Chi Nguyen, M. X. N., & Chi, D. N. (2021). Reflective learning practice for promoting adolescent EFL learners' attention to form. *Innovation in Language Learning and Teaching*, *15*(3), 247-262.

Decristan, J., Hondrich, A. L., Büttner, G., Hertel, S., Klieme, E., Kunter, M., ... & Hardy, I. (2015). Impact of additional guidance in science education on primary students' conceptual understanding. *The Journal of Educational Research*, *108*(5), 358-370.

Fidalgo, R., Torrance, M., Rijlaarsdam, G., Van den Bergh, H., & Álvarez, M. L. (2015). Strategy-focused writing instruction: Just observing and reflecting on a model benefits 6th grade students. *Contemporary Educational Psychology*, *41*, 37-50.

Förster, N., Kawohl, E., & Souvignier, E. (2018). Short-and long-term effects of assessment-based differentiated reading instruction in general education on reading fluency and reading comprehension. *Learning and Instruction*, *56*, 98-109.

Fyfe, E. R., & Rittle-Johnson, B. (2016). The benefits of computer-generated feedback for mathematics problem solving. *Journal of Experimental Child Psychology*, *147*, 140-151.

Gaintza, Z., & Goikoetxea, E. (2016). Spelling instruction in Spanish: a comparison of self-correction, visual imagery and copying. *Journal of Research in Reading*, *39*(4), 428-447.

Gamlem, S. M., Kvinge, L. M., Smith, K., & Engelsen, K. S. (2019). Developing teachers' responsive pedagogy in mathematics, does it lead to short-term effects on student learning?. *Cogent Education*, *6*(1), 1676568.

Gedilki, H., & Buldur, S. (2022). The Effects of Formative Assessment Practices in Science Education on Students' Metacognitive Knowledge and Regulation Skills. *Hacettepe University Journal of Education*, *37*(4), 1393-1415.

Gijlers, H., Weinberger, A., van Dijk, A. M., Bollen, L., & van Joolingen, W. (2013). Collaborative drawing on a shared digital canvas in elementary science education: The effects of script and task awareness support. *International Journal of Computer-Supported Collaborative Learning*, *8*, 427-453.

Hier, B. O., & Eckert, T. L. (2014). Evaluating elementary-aged students' abilities to generalize and maintain fluency gains of a performance feedback writing intervention. *School Psychology Quarterly, 29*(4), 488.

Kegel, C. A., & Bus, A. G. (2012). Online tutoring as a pivotal quality of web-based early literacy programs. *Journal of Educational Psychology, 104*(1), 182.

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

Luquin, M., & Garćia Mayo, M. D. P. (2021). Exploring the use of models as a written corrective feedback technique among EFL children. *System, 98*, 102465.

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.

Nord, A., Hult, H., Kreitz-Sandberg, S., Herlitz, J., Svensson, L., & Nilsson, L. (2017). Effect of two additional interventions, test and reflection, added to standard cardiopulmonary resuscitation training on seventh grade students' practical skills and willingness to act: a cluster randomised trial. *BMJ Open, 7*(6), e014230.

Prinsen, F. R., Terwel, J., Zijlstra, B. J., & Volman, M. M. (2013). The effects of guided elaboration in a CSCL programme on the learning outcomes of primary school students from Dutch and immigrant families. *Educational Research and Evaluation, 19*(1), 39-57.

Puklavec, A., Antekolović, L., & Mikulić, P. (2021). Acquisition of the long jump skill using varying feedback. *Croatian Journal of Education: Hrvatski časopis za odgoj i obrazovanje, 23*(1), 107-132.

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction, 21*(2), 267-280.

Simmons, D. C., Kim, M., Kwok, O. M., Coyne, M. D., Simmons, L. E., Oslund, E., ... & Rawlinson, D. A. (2015). Examining the effects of linking student performance and progression in a Tier 2 kindergarten reading intervention. *Journal of Learning Disabilities, 48*(3), 255-270.

Sukhram, D., & Ellen Monda-Amaya, L. (2017). The effects of oral repeated reading with and without corrective feedback on middle school struggling readers. *British Journal of Special Education, 44*(1), 95-111.

Tavşanlı, Ö. F., Bilgin, A., Yıldırım, K., Rasinski, T., & Tschantz, B. (2021). The effect of a PBWMIP on writing success and attitude toward writing. *Reading and Writing Quarterly*, *37*(5), 425-443.

Truckenmiller, A. J., Eckert, T. L., Codding, R. S., & Petscher, Y. (2014). Evaluating the impact of feedback on elementary aged students' fluency growth in written expression: A randomized controlled trial. *Journal of School Psychology*, *52*(6), 531-548.

Vidal-Abarca, E., Gilabert, R., Ferrer, A., Ávila, V., Martínez, T., Mañá, A., ... & Serrano, M. Á. (2014). TuinLEC, an intelligent tutoring system to improve reading literacy skills/TuinLEC, un tutor inteligente para mejorar la competencia lectora. *Infancia y Aprendizaje*, *37*(1), 25-56.

Wong, H. M. (2017). Implementing self-assessment in Singapore primary schools: Effects on students' perceptions of self-assessment. *Pedagogies: An International Journal*, *12*(4), 391-409.

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.