

Using Formative Assessment to Identify and Address Mathematics Learning Gaps in Grades 6–8 Instruction

¹Chisom Gabrielle Ezechukwu

¹*St. Bernadette Parish School, Burien, Washington, USA*

Email : gabbyeze333@gmail.com

Orcid ID: <https://orcid.org/0009-0001-3612-8351>

Abstract

Mathematics learning gaps in the middle grades are a major barrier to students' academic achievement. This paper presents a quantitative synthesis of the background regarding the use of formative assessments to identify and remediate mathematics learning gaps in Grades 6-8. The synthesis shows that formative assessments improve students' mathematics skills. For example, the sample-weighted effect sizes of the interventions that were studied ranged from 0.16 to 0.52, with a weighted mean of 0.28. Criterion-related validity coefficients between the formative assessments and the state-wide mathematics assessments ranged from 0.61 to 0.80. The receiver operating characteristic (ROC) curves also showed an area under the curve of 0.82, indicating high classification accuracy for students based on their mathematics skills. Four tables in the paper present the results of the literature synthesis on the use of formative assessments in mathematics classrooms. These tables show that the most common learning gaps in mathematics in Grades 6-8 relate to fractions, place value, proportional reasoning, and pre-algebraic thinking. Overall, the results show that using formative assessments in mathematics classrooms is a workable way to reduce learning gaps among students in Grades 6-8.

Keywords: formative assessment, mathematics learning gaps, curriculum-based measurement, Grades 6–8, progress monitoring, effect size, predictive validity

Introduction

The transition from elementary to middle school mathematics is one of the most important events in a student's educational career. In Grade 6 and beyond, the mathematics curriculum shifts away from arithmetic operations towards algebraic concepts, geometric measurements and statistical analyses. Students must have a solid understanding of fractions, ratios, integers and place values prior to entering middle school (National Mathematics Advisory Panel [NMAP], 2008). Students who enter middle school with weak foundations in these areas will experience a series of deficits in their mathematical education that will become increasingly apparent with each passing grade. According to the 2022 National Assessment of Educational Progress (NAEP), only 26% of eighth-grade students scored at or above the Proficient level in mathematics, a decline from pre-2020 assessments of middle school students (National Center for Education Statistics, 2022). However, these statistics do not account for the variability of individual students within a given classroom. Only through the implementation of well-designed assessments can teachers identify these differences in student learning.

Formative assessment, defined as the use of evidence of student learning to adjust instruction prior to or during the instructional process (Black & Wiliam, 1998), is the most direct tool to address the variability of student learning. Unlike summative assessments, which evaluate learning after the instructional sequence has ended, formative assessments happen within the instructional sequence, providing data that can inform adjustments to instruction while it is still possible to make such adjustments. Quantitative research on formative assessment in mathematics has grown significantly over the past two decades, encompassing research on curriculum-based measurement (CBM), technology-assisted diagnostic tools and learning path-aligned diagnostic tools (Foegen, 2008; Confrey et al., 2017; Koedinger et al., 2010). However, no quantitative review of the research on these tools exists to provide an account of their effectiveness in identifying learning gaps in students of Grades 6-8 or of the achievement gains that can result from their use. So, the following research questions are posed to address this gap in the literature:

RQ1. To what extent do formative assessment practices in Grades 6 - 8 mathematics exhibit measurable effect sizes in the improvement of student achievement?

RQ2. How accurately do formative assessment tools predict the performance of middle school students on high-stakes summative mathematics assessments?

RQ3. What quantitative evidence exists regarding the prevalence and pattern of learning gaps in mathematics that are detected by formative assessment tools in Grades 6 - 8?

Research Background

Formative Assessment: Empirical Foundations

The scientific case for formative assessment is based on the work of Black and Wiliam (1998), who conducted a meta-analysis of 250 studies of the impact of formative assessments on student learning

outcomes. The study found that implementing formative assessments increased the average student's learning outcomes by 0.40 to 0.70 standard deviation units, moving the average student from the 50th percentile of the population to the 65th to 73rd percentile. Although these findings were published across a variety of subjects and grade levels, subsequent studies have focused on the effectiveness of formative assessments in mathematics classes, particularly in the middle grades, because gaps in prerequisite knowledge in these classes can be easily identified and remedied through formative assessments. For example, Hattie and Timperley (2007) published a study that explained the relationship between formative assessments and learning outcomes, identifying feedback as the key component of formative assessments. In their study of 196 meta-analyses of educational interventions, they found that feedback had the highest impact on student learning outcomes, with an effect size of $d = 0.73$. Also, their study found that the type of feedback that was provided to students had a important impact upon the effectiveness of the formative assessments, with process-level feedback (feedback that explains why a student made a mistake and provides guidance on how to correct it) having the largest impact upon student learning outcomes.

Curriculum-Based Measurement in Middle School Mathematics

Curriculum-based measurement (CBM) is one of the most rigorously validated implementations of formative assessment in academic domains. Stecker and Fuchs (2000) performed a 20-week study with 42 students of various grade levels (Grades 2-8). Students who received individualized instruction based upon their progress in CBM assessments outperformed students who did not receive such instruction. The study established the logic of implementing data-driven instructional adjustments. Foegen (2008) performed a study with 563 students of various grade levels (Grades 6-8) at four measurement occasions with probes of various types (basic fact fluency, estimation and pre-algebra probes). The estimation and pre-algebra probes exhibited acceptable alternate-form reliability and criterion-related validity, showing that brief probes of 2-4 minutes could yield defensible diagnostic data in mathematics for students of middle school grade levels. The development of the easyCBM measures for middle school mathematics aligned with the Common Core State Standards (CCSS) (Anderson et al., 2012) utilized Item Response Theory (IRT) to create a vertical scale for grades 6-8. The resulting measures comprise 13 alternate forms (10 for weekly measures and 3 for seasonal measures), each of which was established to be equivalent through test characteristic curves (TCCs) and test information functions (TIFs).

Strait et al. (2018) investigated the reliability and validity of aggregated M-CBMs for 283 sixth-grade students. The results indicated that the alternate-form reliability coefficient was 0.85 and that the CBMs were able to accurately identify at-risk students, as well as monitor the progress of those students over time. Also, the study revealed that the CBMs had a high degree of predictive validity regarding the students' academic performance on standardized tests. Also, Helwig et al. (2002) conducted a study of 171 eighth-grade students, of whom 90 were general education students and 81 had learning disabilities. The

researchers found that there was a connection of 0.80 between the 11-item concept-grounded CBM and the California Achievement Test (CAT) scores of the general education students and a connection of 0.61 between the CBM and CAT scores of the students with learning disabilities. Finally, Coddington et al. (2016) conducted a study of 408 students in Grades 6-8. The researchers found that the concepts/application CBM scores of these students correlated with their scores on the CCSS-aligned assessments that were administered to them after the CBM assessments were completed. These findings show that the skills measured by the basic fact CBM were predictive of the skills measured by the concepts/application CBM, which in turn were predictive of the students' CCSS-aligned assessment scores.

Johnson et al. (2013) used logistic regression and ROC analysis to evaluate the relationship between CBM slope measures and performance on state assessments for students in Grades 7-10. The results indicated that the rate-of-improvement slopes derived from CBMs were significantly associated with student performance on state assessments; the AUC value for the model was 0.82, showing that the model was able to classify students as proficient or not proficient with an accuracy similar to that of clinical-grade assessment tools.

Technology-Assisted Formative Assessment Systems

Technology-based platforms allow students to take formative assessment continuously on a large scale. Koedinger et al. (2010) conducted a quasi-experimental study of 1,240 seventh-grade students that utilized the ASSISTments online formative assessment and tutoring platform. The results showed that students who used the ASSISTments platform significantly outperformed students who did not use it on their year-end assessments. The effect sizes for students with special education needs and for students who utilized the platform more frequently were significantly larger than the effect sizes for other students. Roschelle et al. (2016) conducted a randomized controlled trial of 2,850 seventh-grade students from 64 schools who used online mathematics homework with formative feedback. The results indicated that students who completed the online mathematics homework with formative feedback significantly outperformed students who did not complete it on assessments of mathematics skills. The study's effect size was approximately $d = 0.20$. Phelan et al. (2009) studied the technical quality of brief formative assessments taken by sixth-grade students in pre-algebra concepts. The study found that brief formative assessments (4-6 items) were reliable and provided educationally meaningful information about students' understanding of the concepts, particularly in the domains of rational numbers, proportionality, and early algebraic reasoning.

Diagnostic Profiling and Learning Trajectories

Beyond monitoring the progress of students with mathematics learning disabilities, there are various diagnostic tools that have been created to assess the knowledge of individual students. For instance, Confrey et al. (2017) created Math-Mapper 6 - 8, a diagnostic tool that maps the mathematics knowledge of students to the validated learning trajectories for those students. These maps allow teachers to determine which

concepts students struggle with the most, rather than relying upon their overall mathematics knowledge. Three studies have utilized Math-Mapper 6 - 8 to assess the impact of the tool upon the curriculum and the performance of students in early algebra. Also, Yang et al. (2020) created a mental-diagnostic system for junior high school mathematics students based upon a knowledge-and-ability structure. Student feedback regarding the diagnostic tool indicated that over 87% of the students felt that the tool accurately reflected their knowledge of mathematics.

Theoretical Framework

The present analysis utilizes three theoretical constructs: the instructional feedback loop (Black & Wiliam, 1998), CBM theory (Stecker & Fuchs, 2000) and the learning path structure (Confrey et al., 2017).

The Instructional Feedback Loop

The instructional feedback loop is a means of performing formative assessments within the classroom. The formula for this process is as follows: $\Delta I_t = f(T - P_t)$ where ΔI_t represents the adjustments that are to be made to the instruction of the students at time t , T represents the target level of skill for the students and P_t represents the level of skill that is demonstrated by the students at time t . If $P_t < T$, then the students require extra time to reach the target level of skill. If $P_t \geq T$, then the students have reached the target level of skill and can proceed to the next topic in the curriculum. The efficiency of this feedback loop depends upon the frequency with which the students are tested for their level of skill, the accuracy of those measurements of the students' levels of skill and the responsiveness of the teacher to the results of the feedback loop.

CBM Rate of Improvement

Within the CBM structure, formative assessments not only provide a measure of student performance at the time of the assessment but also of the rate of improvement (ROI) that the student exhibits over the period of the assessment. The ROI is calculated using ordinary least-squares regression of the scores obtained from each assessment with the measurement occasion. The ROI can be calculated using the following formula: $ROI = \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$, where x_i represents the measurement occasion (time) and y_i represents the probe score at occasion i . If the calculated ROI falls below a certain threshold, it shows that the instructional approach being implemented is not sufficient to lead to growth in students' skills and so, a modification to the program is required (Johnson et al., 2013; Stecker & Fuchs, 2000).

Learning Trajectory Prerequisites

The learning path structure (Confrey et al., 2017) allows for the organization of the content that will be used to interpret the diagnostic test results. Each skill node within the learning path has a set of prerequisite skills associated with that skill; for example, a node labeled as "Skill Sk " may have a set of prerequisite skills labeled as "Pre(Sk) = { Sk_1 , Sk_2 , ..., Sk_m }". If a student answers incorrectly to an item that is associated

with that skill, then the student is likely to not have mastered any of those prerequisites. So, the goal of the diagnostic is to find the prerequisite skill(s) that the student has not yet mastered and to provide instructional resources targeted towards gaining mastery over that prerequisite skill: Remediation Target = $\underset{M}{\subseq \text{Pre}(S_k)}{\arg\min} |M| \text{ such that } P(\text{mastery of } S_k \mid M \text{ mastered}) \geq \theta$ where θ is a mastery probability threshold (typically $\theta=0.80$ in criterion-referenced frameworks). This target for remediation is mathematically precise, unlike the broad indication of being below grade level.

Research Design and Methodology

Quantitative Synthesis Design

This study uses a quantitative synthesis design that involves collecting and analyzing empirical studies, technical reports and dissertations that meet specific inclusion criteria. The studies are then coded to extract specific quantitative outcomes, which are then synthesized into a complete evidence base. The statistical measure used in this study is the standardized effect size (Cohen's d), which allows for the comparison of the outcomes of different studies with different outcome measures.

Inclusion Criteria

The studies that were to be included in this review had to meet four specific criteria: (a) the study had to include students in Grades 6-8 who were learning mathematics; (b) the study had to include some form of assessment of those students, such as CBM, diagnostic probes, technology-based assessments or embedded feedback; (c) the study had to report on some form of quantitative outcome measure for those students; and (d) the study was published in 2023.

Effect Size Computation

For studies that reported means and standard deviations for two independent groups, Cohen's d was calculated using the following formula: $d = \frac{\bar{X}_1 - \bar{X}_2}{SD_{\text{pooled}}}$ where the pooled standard deviation is: $SD_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}}$. For pre/post single-group designs without a control, the effect size was calculated as: $d = \frac{\bar{X}_{\text{post}} - \bar{X}_{\text{pre}}}{SD_{\text{pre}}}$. A sample-weighted mean effect size across studies was computed as: $\bar{d} = \frac{\sum_{j=1}^k n_j \cdot d_j}{\sum_{j=1}^k n_j}$ where n_j is the total sample size for study j and d_j is its associated effect size.

Predictive Validity Analysis

For studies that report Pearson correlations (r) between formative assessment scores and criterion variables (e.g., statewide test performance), the standard error of estimate (SEE) can be calculated to show the precision of each student's predicted score on the criterion variable. $SE_E = SD_Y \sqrt{1 - r^2}$, where SD_Y is the standard deviation of the criterion variable. The approximate 95% prediction interval for individual predicted scores is $\pm 1.96 \times SEE$.

Results

Effect Sizes Across Formative Assessment Studies

Table 1 presents a synthesis of effect sizes from quantitative studies of formative assessment on student learning in mathematics for Grades 6-8. Each study is categorized by study design type and the effect sizes are reported in Cohen's d units.

Table 1

Effect Sizes From Quantitative Studies on Formative Assessment in Grades 6–8 Mathematics

Study	Grade	N	Design	Formative Tool	Cohen's d
Koedinger et al. (2010)	7	1,240	Quasi-experimental	ASSISTments online platform	0.52
Roschelle et al. (2016)	7	2,850	Randomized (RCT)	Online homework + feedback	0.20
Stecker & Fuchs (2000)	2–8	42	Longitudinal matched	CBM-individualized instruction	0.43
Scafide (2004)	8	620	Quasi-experimental	I Can Learn platform	0.16
Vladescu (2023)	6–8	101	Correlational	Mastery learning cycles	0.18
Weighted mean (Eq. 7)	—	4,853	—	—	0.28

Note. = statistically important at $p < .05$. Effect sizes for Koedinger et al. (2010) and Stecker & Fuchs (2000) are calculated from the means, standard deviations and importance levels reported in those studies using Equations 4 - 6. The effect size for Vladescu (2023) is not statistically important in the Grade 6-8 subsample. RCT = randomized controlled trial; CBM = curriculum-based measurement.

Applying Equation 7:

$$d^- = \frac{(1240)(0.52) + (2850)(0.20) + (42)(0.43) + (620)(0.16) + (101)(0.18)}{1240 + 2850 + 42 + 620 + 101} = \frac{644.8 + 570.0 + 18.1 + 99.2 + 18.2}{4853} = \frac{1350.3}{4853} \approx 0.278$$

The sample-weighted mean effect size of $d^- \approx 0.28$ translates to an approximate gain of 11 percentile points for an average student (from the 50th to approximately the 61st percentile), which is within the range of Hattie and Timperley's (2007) benchmark for process-level feedback interventions in mathematics instruction.

Prevalence of Mathematics Learning Gaps

Table 2 presents the most frequently documented mathematics learning gap domains in Grades 6 - 8, as identified by the diagnostic instruments used in the studies cited. Where a study reported an exact

percentage of students who performed below a specific criterion score, that percentage is presented in the table. Where a study described the prevalence of a learning gap field in more qualitative terms, that description is presented in the table.

Table 2

Approximate Prevalence of Mathematics Learning Gaps Identified by Formative Diagnostic Assessments, Grades 6–8

Skill Domain	Grade Level	% Below Benchmark	Primary Source
Place value and number sense	6	~78%	Huf et al. (2019)
Fraction operations (division, equivalence)	6	~61%	Huf et al. (2019)
Plane geometry and measurement	6	High prevalence†	Huf et al. (2019)
Proportional reasoning/ratio	6–7	~55%‡	Confrey et al. (2017)
Pre-algebraic and conceptual reasoning	7–8	~52%‡	Codding et al. (2016)
Basic arithmetic fact fluency	6–8	~43%‡	Foegen (2008)
Integer operations and signed numbers	7	~48%‡	Phelan et al. (2009)

Note. = qualitative prevalence reported without an exact percentage. ‡ = estimated from reported CBM below-benchmark classification rates or study-reported prevalence patterns; not directly stated as a precise percentage by the cited source. Percentages represent the approximate proportion of students scoring below established formative assessment benchmark criteria. The data in Table 2 reveal a systematic, grade-stratified pattern. The heaviest concentration of learning gaps happens in place value, fraction operations and measurement at Grade 6, which are the basic skills required for proportional reasoning and pre-algebraic thinking in Grades 7 and 8. The relationship between these basic skills and the upper-grade skills is described by Equation 3: the failure to master the skills in Pre(Sk)Pre(Sk) will prevent students from mastering the skills in SkSk without specific remedial instruction.

Predictive Validity of Formative Assessment Instruments

Table 3 presents criterion-related validity and reliability coefficients for the major formative assessment instruments examined in the synthesized literature.

Table 3

Psychometric Characteristics of Formative Assessment Instruments for Middle School Mathematics (Grades 6–8)

Instrument	Grade	N	Criterion Validity (r)	Reliability (r)	Reliability Type	Source
------------	-------	---	------------------------	-----------------	------------------	--------

Concept-grounded CBM (11 items)	8	90 (gen. ed.)	.80	.78	Test-retest	Helwig et al. (2002)
Concept-grounded CBM (11 items)	8	81 (LD)	.61	.72	Test-retest	Helwig et al. (2002)
Estimation probe CBM	6–8	563	.69	.82	Alternate-form	Foegen (2008)
Aggregated random M-CBM	6	283	.74	.85	Alternate-form	Strait et al. (2018)
easyCBM CCSS (IRT-scaled)	6–8	—	.71*	.88	Alternate-form	Anderson et al. (2012)
Algebra CBM (ROC / AUC)	7–10	—	AUC = .82	—	—	Johnson et al. (2013)

Note. LD = Learning disabilities; M-CBM = mathematics curriculum-based measurement; IRT = item response theory; AUC = area under the receiver operating characteristic curve; = estimated from IRT test information functions. Criterion validity is represented as Pearson's r between scores and the state's mathematics assessment, except for Johnson et al. (2013) who used AUC from logistic regression analysis. Using Equation 8 and the data from Helwig et al. (2002), in which $SDY \approx 15$ on the standardized criterion scale:

General education subsample ($r = .80$): $SEE = 151 - (0.80)^2 = 151 - 0.64 = 150.36 = 15 \times 0.600 = 9.0$ points
 $SEE = 151 - (0.80)^2 = 151 - 0.64 = 150.36 = 15 \times 0.600 = 9.0$ points. The 95% prediction interval is:
 $Y \pm (1.96)(9.0) = Y \pm 17.6$ score points
 $Y \pm (1.96)(9.0) = Y \pm 17.6$ score points

Learning disabilities subsample ($r = .61$):

$SEE = 151 - (0.61)^2 = 151 - 0.372 = 150.628 = 15 \times 0.793 = 11.9$ points
 $SEE = 151 - (0.61)^2 = 151 - 0.372 = 150.628 = 15 \times 0.793 = 11.9$ points

The 95% prediction interval widens to: $Y \pm (1.96)(11.9) = Y \pm 23.3$ score points
 $Y \pm (1.96)(11.9) = Y \pm 23.3$ score points

The difference in SEE values between the LD and non-LD populations (9.0 vs. 11.9 points) provides a quantification of the reduced predictive precision for the LD population, justifying the need for supplementary diagnostic instruments (e.g., error analysis, mental-diagnostic profiling) when making high-stakes placement decisions for this population. Also, the AUC of 0.82 in Johnson et al. (2013) provides an

more measure of classification accuracy. The AUC is the probability that a randomly selected proficient student will receive a higher CBM score than a randomly selected non-proficient student. So, an AUC of 0.82 shows that the algebra CBM correctly rank-orders proficient and non-proficient students approximately 82% of the time, which is significantly above the 0.50 chance-level baseline. Finally, the r^2 coefficient of determination across the instruments in Table 3 ranges from $r^2=(0.61)^2=0.37$ to $r^2=(0.80)^2=0.64$. So, well-designed middle-grades formative probes explain 37% to 64% of the variance in end-of-year statewide mathematics performance for those students.

Instructional Decision Thresholds

The second category of outcomes from the implementation of formative assessment systems are the decision thresholds that have been established for each tier of instruction within the classroom. These thresholds were established through empirical analysis of student performance and have been published in various studies on the topic. Table 4 presents some of these thresholds as established by those studies.

Table 4

Formative Assessment Decision Thresholds for Mathematics Instruction in Grades 6–8

Grade	Benchmark Instrument	On Track (Score $\geq$)	At Risk (Score Range)	Intensive Need (Score $\leq$)	ROI Criterion (pts/week)	Source
6	M-CBM (aggregated random)	≥ 40	25–39	≤ 24	≥ 1.5	Strait et al. (2018)
7	Algebra CBM	≥ 55	35–54	≤ 34	≥ 1.2	Johnson et al. (2013)
8	Concept-grounded CBM	≥ 70	50–69	≤ 49	≥ 1.0	Helwig et al. (2002)
6–8	easyCBM CCSS (IRT-scaled)	$\geq 65^{\text{th}}$ %ile	35 th –64 th %ile	$\leq 34^{\text{th}}$ %ile	—	Anderson et al. (2012)

Note. The scores are presented in the native metric of each of the instruments. The ROI is calculated using the OLS regression slope of Equation 2. The percentiles are calculated using the IRT-calibrated national norming samples for easyCBM. Students with an ROI that falls below the criterion for three consecutive weeks will trigger a decision to change the instructional program in which the student is participating. The three-tiered system is directly related to the Response to Intervention (RTI) structure, which assigns students to one of three tiers of instruction: Tier 1 (universal classroom instruction), Tier 2 (targeted small-group supplemental instruction) and Tier 3 (intensive individualized instruction). The ROI criterion for three consecutive weeks is related to Equation 1: if a student's growth slope falls below the criterion value,

then the student's PtPt path will not reach T within the available instructional window, triggering the decision function $\Delta It > 0$ and a program modification.

Discussion

Interpretation of Effect Sizes

The sample-weighted mean effect size of $d \approx 0.28$ falls within the small-to-medium range according to Cohen's (1988) criteria for effect sizes, where $d = 0.20$ is considered small and $d = 0.50$ is medium. Each of these factors contributes to the attenuation of the effect size of the intervention relative to those that have been conducted within laboratory settings. However, when these variables are controlled for, such as in Koedinger et al.'s (2010) study, the effect size of the intervention was found to be $d = 0.52$, which is considered to be practically important. Also, the results of Koedinger et al. (2010) show that the dose of the intervention had an impact upon the learning outcomes of the students who participated in the intervention. Specifically, the results show that the more frequently that formative assessments were administered to the students, the more likely that those students would exhibit learning gains as a result of the intervention. These findings are consistent with Equation 1, which shows that the more frequently that PtPt is measured and ΔIt is adjusted for the students, the closer that the instructional methods will track the learning trajectories of each individual student and the less likely that a gap between the students' learning and the instruction will accumulate prior to being detected and addressed.

Diagnostic Precision and Learning Disabilities

The differential validity coefficients published by Helwig et al. (2002) were 0.80 for general education students and 0.61 for students with learning disabilities. The increased standard error of measurement for students with learning disabilities from 9.0 to 11.9 means that the confidence interval for predictions based on CBM measures will be 12 points wider on each side for students with learning disabilities compared to general education students. So, teachers should not use CBM results alone to make important educational decisions about students with learning disabilities. Instead, teachers could use the results from a mental-diagnostic assessment to supplement their decisions. For instance, Yang et al. (2020) developed a mental-diagnostic assessment that measured students' knowledge of various mathematical concepts and found that the assessment was endorsed by 87% of students who took the assessment. Also, the assessment is able to measure students' knowledge of concepts in a way that allows teachers to recognize which concepts are understood by the students and which concepts may need to be taught to those students.

The Sequential Structure of Middle-Grades Learning Gaps

The pattern in Table 2 reflects the hierarchical structure of middle-grades mathematics knowledge. The heaviest loading of learning gaps at Grade 6 - particularly in place value ($\approx 78\%$), fraction operations ($\approx 61\%$) and geometric measurement - is not incidental; these are precisely the domains that the NMAP (2008) identified as basic prerequisites for algebra readiness. When these gaps are not detected and addressed

within Grade 6, they propagate into Grades 7 and 8 as proportional reasoning and pre-algebraic deficits, as evidenced by the Coddington et al. (2016) and Phelan et al. (2009) findings. So, early formative identification of these knowledge gaps at Grade 6 offers high instructional use. The learning path structure (Confrey et al., 2017) formalizes this logic through Equation 3: the minimal prerequisite remediation set for restoring a student's access to on-grade content in Grades 7 - 8 is determined by tracing backward through the dependency graph to the earliest unmastered nodes - nodes that, in the majority of at-risk students, reside in the fraction and place-value domains introduced or consolidated in Grades 4 - 6.

Technology-Mediated Feedback and Process-Level Intervention

The relatively larger effect size observed in the ASSISTments study ($d = 0.52$; Koedinger et al., 2010) compared to paper-based CBM studies can be attributed to the quality of feedback provided by technology-assisted platforms. Hattie and Timperley (2007) found that process-level feedback (i.e., explaining why an error occurred and providing a corrective strategy) produces substantially larger effects than score reporting alone. Technology-assisted platforms are able to provide this type of feedback automatically to students at the point of error and at a per-student level that would be impractical for a teacher to provide in a classroom of 25-30 students. Also, Wiliam (2011) stated that the value of formative assessments is only realized if the assessment information is used by both the teacher and the student. Platforms like ASSISTments and the Math-Mapper system (Confrey et al., 2017) automatically use the diagnostic data from the students to inform the learning tasks that are presented to them, eliminating the need for teachers to implement these assessments and reducing the latency between the assessment and the response from the student.

Limitations

Several limitations exist regarding the scope of the conclusions that can be drawn from this analysis. First, the effect sizes that are calculated for each of the studies included in Table 1 are based upon different types of outcome measures. These outcome measures are not necessarily comparable to one another, even after standardization of those measures. Second, the majority of the studies that were included in this analysis were conducted within the setting of public schools within the United States. So, the generalizability of the findings from these studies to other countries or educational systems is limited. Third, the decision thresholds that are presented in Table 4 were validated within specific contexts and populations. So, any attempts to apply those thresholds to other populations or settings would require the use of local data to calibrate those thresholds to the specific populations that are to be assessed. Finally, the role of teacher professional development in moderating the effectiveness of formative assessments is a variable that was discussed in several of the studies that were reviewed for this analysis (Wiliam, 2011). However, that variable was not quantitatively addressed in this analysis.

Conclusion

This study aimed to determine whether formative assessment can help identify learning gaps in mathematics among students in Grades 6-8. The results of this study show that formative assessment can be used to identify learning gaps in mathematics among students in Grades 6-8. More specifically, the results of this study show that effect sizes for formative assessment interventions in middle school mathematics range from $d = 0.16$ to $d = 0.52$, with a sample-weighted mean of $\bar{d} \approx 0.28$. Also, technology-assisted platforms that provide students with feedback on their mathematics practice tend to exhibit the largest effect sizes for mathematics learning gaps among students in Grades 6-8. Also, formative assessments that focus on specific mathematics concepts, such as estimation, conceptual application and algebraic reasoning, have criterion-related validity coefficients ranging from $r = 0.61$ to $r = 0.80$ compared to statewide mathematics assessments, which shows that these formative assessments can account for 37% to 64% of the variance in student performance on those statewide assessments. Finally, the area under the curve (AUC) for algebra-specific classifying behavior matrices (CBMs) is 0.82, which shows that the CBMs can accurately identify students with learning gaps in algebra at the Grade 7-8 level. Overall, the results of this study show that the most common learning gaps among students in Grades 6-8 relate to fractions, place value, proportional reasoning and pre-algebraic thinking. These results can be used to develop decision thresholds for identifying learning gaps in mathematics among students in Grades 6-8, which can then be used to differentiate instruction for those students within Response to Intervention (RTI) frameworks.

The instructional feedback loop (Equation 1) must be closed not once per unit or semester but on a biweekly or weekly basis, with adjustments to instruction based upon the data from the path of student learning. If such a standard can be met within a given educational institution, then the evidence of these studies suggests that the achievement gap in middle-grades mathematics will be reduced.

References

- Anderson, D., Irvin, P. S., Patarapichayatham, C., Alonzo, J., & Tindal, G. (2012). *The development and scaling of the easyCBM CCSS middle school mathematics measures* (Technical Report No. 1207). Behavioral Research and Teaching, University of Oregon.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74. <https://doi.org/10.1080/0969595980050102>
- Codding, R. S., Mercer, S. H., Connell, J. E., Fiorello, C. A., & Kleinert, W. L. (2016). Mapping the relationships among basic facts, concepts and application, and Common Core curriculum-based mathematics measures. *School Psychology Review*, 45(1), 19–38. <https://doi.org/10.17105/SPR45-1.19-38>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

- Confrey, J., Gianopulos, G., McGowan, W., Shah, M., & Belcher, M. (2017). Scaffolding learner-centered curricular coherence using learning maps and diagnostic assessments designed around mathematics learning trajectories. *ZDM Mathematics Education*, 49(5), 717–734. <https://doi.org/10.1007/s11858-017-0869-1>
- Foegen, A. (2008). Progress monitoring in middle school mathematics: Options and issues. *Remedial and Special Education*, 29(4), 195–207. <https://doi.org/10.1177/0741932507309716>
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Helwig, R., Anderson, L., & Tindal, G. (2002). Using a concept-grounded, curriculum-based measure in mathematics to predict statewide test scores for middle school students with LD. *Journal of Special Education*, 36(2), 102–112. <https://doi.org/10.1177/00224669020360020501>
- Huf, V. B. S., Huf, S. F., Pinheiro, N. A. M., & Burak, D. (2019). Avaliação diagnóstica no 6º ano: O que ela mostra em relação aos conteúdos matemáticos [Diagnostic evaluation in the 6th year: What it reveals about mathematical content]. *Brazilian Journal of Development*, 5(12), 30600–30613. <https://doi.org/10.34117/BJDV5N12-176>
- Johnson, E. S., Galow, P. A., & Allenger, R. (2013). Application of algebra curriculum-based measurements for decision making in middle and high school. *Assessment for Effective Intervention*, 39(1), 3–11. <https://doi.org/10.1177/1534508412461435>
- 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. <https://doi.org/10.2190/EC.43.4.D>
- National Center for Education Statistics. (2022). *NAEP 2022 mathematics: National results overview*. U.S. Department of Education. <https://nces.ed.gov/nationsreportcard/mathematics/>
- National Mathematics Advisory Panel. (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel*. U.S. Department of Education.
- Phelan, J., Kang, T., Niemi, D., Vendlinski, T. P., & Choi, K. (2009). *Some aspects of the technical quality of formative assessments in middle school mathematics* (CRESST Report No. 750). National Center for Research on Evaluation, Standards, and Student Testing, UCLA.
- Roschelle, J., Feng, M., Murphy, R., & Mason, C. A. (2016). Online mathematics homework increases student achievement. *AERA Open*, 2(4), 1–12. <https://doi.org/10.1177/2332858416673968>
- Scafide, K. (2004). *Effects of I Can Learn on math achievement in Gwinnett County middle school* [Unpublished manuscript]. <https://scispace.com/papers/effects-of-i-can-learn-on-math-achievement-in-gwinnett-wikwx9pnce>

- Stecker, P. M., & Fuchs, L. S. (2000). Effecting superior achievement using curriculum-based measurement: The importance of individual progress monitoring. *Learning Disabilities Research and Practice*, 15(3), 128–134. https://doi.org/10.1207/SLDRP1503_2
- Strait, G. G., Smith, B. H., & McQuillin, S. D. (2018). Aggregated randomly generated math curriculum-based measurements for middle school students: Reliability, predictive validity, and cut score precision. *Assessment for Effective Intervention*, 44(1), 58–64. <https://doi.org/10.1177/1534508418761231>
- Vladescu, C. (2023). The relationship between mastery learning models and academic achievement in mathematics. *International Electronic Journal of Mathematics Education*, 18(4), em0741. <https://doi.org/10.29333/iejme/13705>
- Wiliam, D. (2011). *Embedded formative assessment*. Solution Tree Press.
- Yang, H., Wang, H., Wang, H., Liang, M., & Huang, T. (2020). The cognitive diagnosis based on the framework of junior high school mathematics knowledge and ability. In *Proceedings of the 2020 International Symposium on Educational Technology* (pp. 253–257). IEEE. <https://doi.org/10.1109/ISET49818.2020.00064>