

EXECUTIVE SUMMARY

The purpose of this report is to provide context and process evaluation information related to key components of the Assessment for Learning program. The focus of the 2020-21 evaluation was to provide context and determine the level of implementation of the grant components, perceptions of staff who participated in grant activities, and the impact of the program on student achievement. Data for this report were obtained from interviews with program staff members, a review of internal documents, and analyses of available survey and student achievement data. Dallas ISD used \$186,357 in Title II grant funds to implement the program in 2020-21. The program hired a coordinator to directly support campuses, lead professional development sessions, and coach teachers and campus leaders.

The focus of the 2020-21 evaluation was to determine the level of implementation of the grant-funded programs and to examine participant perceptions of the programs, as applicable. Data for this report were obtained from interviews with program staff members assigned to each grant component, a review of internal documents, and analyses of available survey data.

EVALUATION QUESTIONS/OUTCOMES

How was the Assessment for Learning program implemented?

Program Characteristics

In 2021-21, the Assessment for Learning program took place on eleven campuses: Ben Milam, Kiest, Zaragoza, Stevens Park, Connor, and Tatum Elementary Schools; Franklin and Spence Middle Schools; and Woodrow Wilson, Molina, and Adamson High Schools. A total of 139 teachers, across all grade levels and content areas, implemented the program in their classrooms. Assessment for Learning utilized an innovative and highly communicative approach to teaching that emphasizes formative assessment throughout instruction.

Formative assessment is a planned, ongoing process used by students and teachers during learning and teaching to elicit and use evidence of student learning to improve student understanding of intended content area learning outcomes and support students to become self-directed learners. Effective use of the formative assessment process requires students and teachers to integrate and embed the following practices in a collaborative and respectful classroom environment: clarifying learning goals and success criteria within a broader progression of learning; eliciting and analyzing evidence of student thinking;

engaging in self-assessment and peer feedback; providing actionable feedback; and using evidence and feedback to move learning forward by adjusting learning strategies, goals, or next instructional steps (Fisher, Frey, & Hattie, 2018). Teachers operate in daily feedback loops where students and teachers answer the following questions: Where am I going? Where am I now? How do I close the gap? or Where to next? (Sadler, 1998).

Campus Walks and Observations

Assessment for Learning program staff utilize the Formative Assessment Rating Rubrics and Observation Protocol (FARROP) to determine implementation of the ten dimensions of formative assessment in classrooms. The Assessment for Learning coordinator conducted a total of four campus walks throughout the year (November, February, March, and May). Because the program's results were noted informally and did not involve collecting quantitative measures, an analysis of results was not possible.

In addition, program staff and campus leaders provided coaching and training for teachers in the year three pilot program through observations and POD (Plan, Observe, Debrief) cycles. A POD cycle is a one-and-a-half-day process where teachers plan together, observe each other's classrooms, and receive feedback from one another and their trainer. In previous years, the program conducted formal POD cycles in November, January, and March. However, due to the shortage of substitute teachers in 2020-21, formal POD cycles were impossible. Instead, campuses engaged in "mini-PODs" where they used conference or PLC time to observe each other. At other campuses, teachers recorded video of lessons and used the POD protocol to give each other feedback after viewing.

Teacher Training

The Assessment for Learning program provided several traditional professional development opportunities throughout the year in addition to the ongoing observation and coaching cycles. A total of 1,137 district teachers, principals, and central staff members attended an online presentation by Doug Fisher on September 1, 2020. A total of 2,532 district teachers, principals, and central staff members attended a professional development session led by formative assessment expert John Hattie on January 4, 2021. Further, 4,236 district teachers, principals, and central staff members attended training with John Almarode on improving distance learning, on March 10, 2021. Assessment for Learning teachers also had access to asynchronous digital training materials throughout the year.

via Schoology, a learning management system managed by PowerSchool.

Simultaneous Learning

Teachers within the Assessment for Learning program implement simultaneous learning, with distance and in-person students in the same classroom at the same time. Through this approach, the program seeks to create an equitable environment through shared culture, access, and expectations for both in-person and distance learners. Teachers were trained to design the physical classroom environment to support both distance and in-person learners while selecting appropriate videoconferencing tools to leverage screen sharing for equitable access. Teachers established norms and expectations for both cohorts of students without attempting to plan two different courses.

Simultaneous Learning emphasized providing clear learning goals, success criteria, and expectations and building in opportunities for collaboration, feedback, and assessment. Teachers explained the learning goals, success criteria, norms, and expectations. They emphasized assigning student roles to promote agency and collaboration, while building in opportunities for teacher- and student-led small groups. Teachers were encouraged to intentionally create moments of relationship-building within and across modalities of in-person and distance learning.

What were teacher perceptions of the Assessment for Learning program?

Methodology

The Assessment for Learning program utilized Google Forms to craft and administer a feedback survey exploring teachers' evolving implementation of Assessment for Learning practices. The program manager provided an online survey link to Assessment for Learning teachers at three different points throughout the year. The beginning-of-year (BOY) survey link was open from September 10, 2020 to November 17, 2020. The middle-of-year (MOY) survey link was open from January 23, 2021 to March 7, 2021. The end-of-year (EOY) survey link was open from May 24, 2021 until June 18, 2021. The purpose of the survey was to collect feedback related to participants' individual experiences working in the Assessment for Learning program. The survey included one demographic item and 10 items with Likert-type response options rating staff perceptions of the program. All Assessment for Learning teachers ($n = 124$) were selected to receive the survey. Program staff shared raw survey data with the program evaluator. The evaluator used frequency analyses to determine rates of positive responses for each scaled item, along with the number and percentage of responses per category for other items.

Results

A total of 115 respondents completed at least one survey, resulting in a 93 percent response rate, though participation dropped sharply as the year progressed (61 teachers responded during the MOY survey window, and 12 teachers responded during the EOY survey window.)

Overall, the proportion of teachers who indicated that they incorporated each behavior into their teaching increased over the course of the school year. The largest increases were seen in posting success criteria that tells students what successful learning will look like (61.8% at BOY versus 83.3% at EOY, an increase of 21.5 percentage points) and providing opportunities for students to set goals, reflect, and monitor their own progress (42.6% at BOY versus 75.0% at EOY, an increase of 32.4 percentage points). At the beginning of the year, 76.5 percent of teachers reported that they modeled how to use mistakes, failure, and struggle as opportunities for growth. By the end of the year, 91.7 percent of teachers reported "often" or "usually" doing so. The formative assessment skill rated least implemented at the beginning of the year (28.7% at BOY), teaching students how to give and receive feedback, increased by 13 percentage points by the end of the year (41.7% at EOY).

How did participation in the Assessment for Learning program affect student achievement?

Methodology

The NWEA *Measures of Academic Progress* (MAP) assessments are universal screeners administered to Dallas ISD students in grades one through ten at the beginning (BOY), middle (MOY), and end (EOY) of the school year. MAP data consists of several variables used to describe student performance and academic growth over time by subject. The Rasch Unit (RIT) score represents the level at which students are ready to learn at the time of assessment. The RIT score is a standardized, equal-interval score that can be compared to other RIT scores regardless of student age or grade. The second MAP Growth variable used for this report was the Conditional Growth Percentile (CGP). The CGP represents the extent to which student achievement changed from one testing window to the next, relative to change for other students who are similar in starting achievement level, grade, and number of instructional weeks between testing windows. For each subject assessed, CGPs for each student are calculated with respect to a large, nationally-normed sample obtained by NWEA and can be calculated whenever a student takes a MAP assessment in a new testing window (e.g., BOY to MOY, MOY to EOY, BOY to EOY). A CGP at the 50th percentile indicates that the RIT score increase from the prior testing window represents the median amount, or expected level, of progress for that grade level¹.

The evaluator used student-level MAP data files dated November 5, 2020 (BOY), March 30, 2021 (MOY), and June

¹ For more information, see <https://community.nwea.org/docs/DOC-1630>.

23, 2021 (EOY) and merged them with the district demographic data files pulled closest to the dates of the *MAP Growth* data files for each testing window. The evaluator selected a sample of 50 grades one through 11 Assessment for Learning teachers. Students were flagged as participants if they were identified by district course files as being taught by one of these teachers. To construct comparison groups of comparable classrooms not implementing the program, classrooms of the same grade on the same campus as implementing teachers were chosen whenever possible. In the case of Ben Milam Elementary, where the entire campus was implementing the program, eight teachers from the same grade levels at Lipscomb Elementary were selected. This campus was chosen based on similarity of campus-level demographic and achievement criteria used to create state accountability comparison groups. Students were selected into the sample if they had valid *MAP* scores for all three test administrations (English language only in Reading, as English and Spanish *MAP* Reading assessments are not directly comparable). In the cases where students appeared on rosters for both AfL and non-AfL teachers (such as high school students taking courses from multiple teachers on a campus), the student was omitted from the study. The evaluator then used frequency and crosstabs analyses to calculate the average *MAP* RIT and CGP scores by subject, grade, and student group. The evaluator conducted *t*-tests of mean RIT scores to ascertain if the program made a statistically significant difference in the participants' academic growth as measured by *MAP* results. Also, the evaluator used Cohen's *d* to examine if differences were practically meaningful. Finally, the evaluator conducted chi-square tests of independence to determine if a significant association existed between AfL participation and growth threshold rates as illustrated by the CGP score.

Results

At all grade levels, RIT scores for reading and mathematics generally remained unchanged or increased from BOY to MOY and from MOY to EOY. For each assessment and testing period, mean RIT scores generally increased for higher grade levels, as expected due to the nature of the assessment (scores are given on a single continuum of RIT scores from grades one through 11).

While AfL students' overall mean Reading RIT across all grades was five points higher than non AfL students at BOY (210 versus 205), the mean RIT for AfL students increased by three points by MOY (210 to 213), while increasing by four points for non-AfL students (205 to 209). From MOY to EOY, AfL students across all grades lost the previous gains, with an overall mean RIT score of 210, the same as at BOY. For non-AfL students, the overall mean EOY RIT score was three points higher than at BOY (205 versus 208). Notably, students at grades one through three saw larger gains from BOY to EOY than upper grades: 11 to 18 points for AfL students and nine to 15 points for non-AfL students. In

addition, AfL students at grade eight saw a loss of six points (206 to 200) from BOY to EOY, while non-AfL students gained one point during this time period (227 to 228).

While there were slight differences in mean RIT scores, *t*-test and effect size analyses of the differences between AfL and non-AfL students' mean RIT Reading scores at BOY, MOY, and EOY showed no statistical or practical significance.² For BOY, $t(20) = 0.17$, $p = 0.86$, *Cohen's d* = 0.19. For MOY, $t(20) = 0.22$, $p = 0.82$, *Cohen's d* = 0.18. For EOY, $t(20) = 0.28$, $p = 0.78$, *Cohen's d* = 0.10.

Overall mean Mathematics RIT scores for AfL students was three points higher at BOY than for non-AfL students (212 versus 209). From BOY to MOY, the mean RIT increased by eight points for both AfL students (212 to 220) and non-AfL students (209 to 217). From MOY to EOY, AfL students across all grades gained two points (220 to 222). For non-AfL students, the mean RIT score also increased by three points from MOY to EOY (217 to 220). Again notably, students at lower grades, this time one through three, saw notable gains from BOY to EOY, 15 to 23 points for AfL students and 11 to 16 points for non-AfL students. Again, *t*-test and effect size analyses of the differences between AfL and non-AfL students' mean RIT Mathematics scores at BOY, MOY, and EOY showed no statistical or practical significance. For BOY, $t(20) = 0.37$, $p = 0.72$, *Cohen's d* = 0.13. For MOY, $t(20) = 0.28$, $p = 0.78$, *Cohen's d* = 0.09. For EOY, $t(20) = 0.32$, $p = 0.76$, *Cohen's d* = 0.08.

To measure growth across testing periods, the evaluator divided CGP scores into three categories based on NWEA's definition of meeting expected growth. Students with a CGP below the 45th percentile "failed to meet" nationally-normed growth expectations. Those with a CGP between the 46th and 55th percentile "met" expectations, and those with a CGP above the 55th percentile "exceeded" those expectations for growth.

For reading, the percentage of AfL students who met or exceeded growth thresholds by MOY was 55.4 percent which was lower by 2.5 percentage points compared to Non-AfL students (57.9%). For mathematics, the percentage of AfL students who met or exceeded growth thresholds by MOY was 71.1 percent which was higher by 1.7 percentage points compared to Non-AfL students (69.4%). There were no significant differences between students of AfL and Non-AfL teachers in meeting or exceeding growth expectations (2.5 percentage points or less). Further, a chi-square test of independence showed that there was no significant association between AfL participation and growth threshold rates in either Reading ($X^2 = 5.03$, $p = .08$) or Mathematics ($X^2 = 0.49$, $p = .78$).

For reading, the percentage of AfL students who met or exceeded growth thresholds by EOY was 34.2 percent which was lower by 4.8 percentage points compared to

² A *p*-value less than 0.05 (typically ≤ 0.05) is statistically significant. A *p*-value higher than 0.05 (> 0.05) is not statistically significant. Cohen's *d*

values of .20, .50, and .80 are considered small, medium, and large, respectively (Cohen, 1992).

Non-AfL students (39.0%). For mathematics, the percentage of AfL students who met or exceeded growth thresholds by EOY was 49.4 percent which was lower by 1.5 percentage points compared to Non-AfL students (50.9%). There were no significant differences between students of AfL and Non-AfL teachers in meeting or exceeding growth expectations (4.8 percentage points or less). Further, a chi-square test of independence showed that there was no significant association between AfL participation and growth threshold rates in either Reading ($X^2 = 5.03$, $p = .08$) or Mathematics ($X^2 = 0.26$, $p = .88$).

What are the plans for the future of the Assessment for Learning program?

Methodology

Evaluation of the Assessment for Learning program included interviews with program management and a review of internal documents. The purpose of the interviews was to gather information about program requirements and activities. Follow-up information was collected by email.

Results

The Assessment for Learning Design Team was an advisory group consisting of program management, representatives from the Teaching and Learning and Evaluation and Assessment departments, AfL campus principals, and AfL teachers. The Design Team met four times during the year, in November, December, February, and June, providing feedback on program implementation, sharing campus and classroom challenges and solutions, and advising program management on program expansion plans.

On Saturday June 12, 2021, the program held an Assessment for Learning Summit on Zoom. The event lasted from 9:00 am to 1:00 pm and was open to all AfL teachers and principals. Three guest presenters joined district staff for author- and teacher-led breakout rooms, and discussions around books supplied to attendees upon registration. Formative assessment expert John Almarode addressed learning goals, success criteria, and feedback. Author Anthony Muhammad discussed equity and transforming school culture. Author and educational consultant Dr. Nicole Law presented on culturally

responsive pedagogy and visible learning. A total of 198 district staff members attended.

While the program was implemented by at least some teachers at 11 campuses in 2020-21, the district is seeking to expand the program to include additional high priority campuses in 2021-22. Additionally, the program will expand to campus-wide implementation at several existing AfL campuses including Adamson and Wilson High Schools, Spence Middle School, and Conner, Stevens Park, Tatum, and Zaragoza Elementary Schools. Program management and district leadership are closely monitoring data to see if these practices could potentially be scaled district-wide as early as 2022-23.

RECOMMENDATIONS

Based on the key findings, the following recommendations are offered for stakeholders.

Formalize campus walk feedback. Currently, campus walk data is captured through non-standardized observer notetaking. Formalizing the campus walk feedback process to include an observation checklist would allow for quantitative analysis of the quality of instruction observed, including tracking change over time.

Continue strong collaboration practices, including design team, PODs, and campus walks. Applying new formative assessment approaches across schools is challenging because it aims to fundamentally change the relationship between teachers and students. A formal approach to collaboration and sharing of best practices between teachers and campus administrators increases the chance for success.

REFERENCES

Fisher, D., Frey, N., Hattie, J. (2018). Developing Assessment-Capable Visible Learners, Grades K-12: Maximizing Skill, Will, and Thrill. United States: SAGE Publications.

Sadler, D. Royce (1998) Formative Assessment: Revisiting the Territory, *Assessment in Education: Principles, Policy & Practice*, 5(1), 77-84.