WestEd

An Evaluation of the Air Tutors Program: A 1-Year Follow-Up on Math Achievement

Findings From an Arnold Ventures Grant

Anthony Fong
Min Chen-Gaddini
Min Huang

February 2026

Suggested citation: Fong, A., Chen-Gaddini, M., & Huang, M. (2026). *An evaluation of the Air Tutors program: A 1-year follow-up on math achievement.* WestEd.

WestEd is a nonpartisan, nonprofit organization that aims to improve the lives of children and adults at all ages of learning and development. We do this by addressing challenges in education and human development, increasing opportunity, and helping build communities where all can thrive. WestEd staff conduct and apply research, provide technical assistance, and support professional learning. We work with early learning educators, classroom teachers, local and state leaders, and policymakers at all levels of government.

For more information, visit WestEd.org. For regular updates on research, free resources, solutions, and job postings from WestEd, subscribe to the E-Bulletin, our semimonthly e-newsletter, at WestEd.org/subscribe.

Contents

Acknowledgement

This follow-up study would not have been possible without the generous support, dedication, and contributions of numerous individuals and organizations. We extend our sincere gratitude to the following groups and partners:

Arnold Ventures foundation

We deeply appreciate Arnold Ventures' continued financial support and commitment to advancing rigorous educational research. Its investment in both the original evaluation and this extended analysis made this work possible.

Teachers and students from the study districts

We are grateful to the students and teachers who participated in the Air Tutors program and contributed data over multiple years. Their engagement and cooperation provided valuable insights into both the short- and long-term impacts of the program.

Administrators from the study districts

We appreciate the support of district and school administrators in District A and District B. Their coordination, data sharing, and commitment to student learning greatly facilitated the implementation of both the initial evaluation and this long-term follow-up study.

Air Tutors team

We thank the Air Tutors team for their collaboration throughout the evaluation period. Their openness in discussing program implementation and tutoring practices helped deepen our understanding of the intervention.

Executive Summary

This brief is a follow-up to WestEd's 2025 evaluation of the Air Tutors high-dosage math tutoring program (Fong et al., 2025). Whereas the 2025 evaluation assessed the impact of the Air Tutors program on short-term student outcomes, the evaluation described in this brief assessed the impact of the Air Tutors program on long-term student outcomes. More specifically, this evaluation sought to answer the following research question:

> Does student participation in the Air Tutors program have a positive impact on math achievement as measured by the state standardized summative math assessment that is administered 1 year after the delivery of the tutoring?

Methods

Similar to the prior study (Fong et al., 2025), this evaluation employed a quasi-experimental design using Mahalanobis distance matching to compare students who received Air Tutors math instruction (treatment group) with observationally similar students who did not receive tutoring (comparison group). Two analytic samples were examined: a full-sample analysis (i.e., all students who attended at least one Air Tutors session) and a 24-hour dosage analysis (i.e., students who received at least 24 hours of tutoring, which represents approximately 80 percent of the intended 30-hour dosage).

The outcome measure for this brief was students' standardized scale scores (*z*-score by state and grade) on the statewide standardized summative math assessments administered in spring 2025, approximately 1 year after the tutoring intervention concluded in the 2023–24 school year. These statewide assessments were the Colorado Measures of Academic Success (CMAS) in District A and the State of Texas Assessments of Academic Readiness (STAAR) in District B.

Baseline Equivalence

For both the full-sample and the 24-hour dosage analyses, baseline equivalence was achieved across all key student characteristics, including gender, economic disadvantage status, English Learner status, special education status, and prior-year math achievement. Standardized mean differences for all variables were near zero and well below the What Works Clearinghouse threshold of 0.25 standard deviations, confirming that the treatment and comparison groups were statistically equivalent on observables before tutoring.

Key Findings

Full sample analysis: When pooling results across both districts, students who participated in the Air Tutors program scored slightly lower than did comparison students on the spring 2025 statewide math assessment, but the difference was not statistically significant (effect size = –0.035; p = .332). Similar small and not statistically significant differences were found when analyzing District A and District B separately. These results indicate no statistically significant long-term impacts of program participation on statewide math performance for the overall student sample.

24-hour dosage analysis: Among students who received at least 24 hours of tutoring, the combined effect across districts was also small and not statistically significant (effect size = –0.012; p = .804). District-specific estimates were also near zero. The results show that for students meeting the recommended dosage threshold, no measurable long-term effects were detected on the spring 2025 statewide math assessment.

Conclusion: One year after tutoring ended, Air Tutors students did not exhibit measurable differences in statewide math performance relative to matched peers, even among those who completed at least 24 hours of tutoring.

Introduction

As a follow-up to a 2025 report that examined the short-term effects of the Air Tutors high-dosage math tutoring program on student achievement (Fong et al., 2025), this brief examines the long-term impacts of the Air Tutors program. More specifically, the evaluation described in this brief sought to answer the following research question:

> Does student participation in the Air Tutors program have a positive impact on math achievement as measured by the state standardized summative math assessment that is administered 1 year after the delivery of the tutoring?

Long-Term Impact of Air Tutors on Student Achievement

This section presents results of the evaluation of the Air Tutors program's long-term impact on student academic achievement. It begins with a quick review of the evaluation methodology and the outcome measure used in the evaluation. Next, it reviews the results from the baseline equivalence tests, followed by a presentation of the impact findings from both the full-sample analysis and the 24-hour dosage analysis.

Impact Evaluation Methodology

This impact evaluation employed a matching analysis in its analytic design. This analysis entailed analytically matching students who received math tutoring from Air Tutors (treatment group) with observationally similar students who did not receive tutoring from Air Tutors or any other math tutoring program during the 2023–24 school year evaluation period (comparison group). For additional details of the methodology, please refer to the 2025 Air Tutors evaluation, which used the same methodology (Fong et al., 2025).

Outcome Measure

The outcome measure is the statewide standardized summative math assessments for students in grades 4, 5, and 6 during the 2024–25 school year in two school districts: District A in Colorado and District B in Texas.

- In Colorado, the statewide assessment system is known as the Colorado Measures of Academic Success (CMAS). The Colorado Department of Education oversees the administration of CMAS assessments, which are designed to measure students' mastery of the Colorado Academic Standards in various content areas.
- In Texas, the State of Texas Assessments of Academic Readiness (STAAR) is the standardized testing program administered by the Texas Education Agency. STAAR assessments are designed to measure the extent to which students have learned and can apply the knowledge and skills defined in the Texas Essential Knowledge and Skills for each grade level and subject area.

Impact Results

Two separate analyses were conducted during the evaluation of the long-term impact of the Air Tutors program: (a) a full-sample analysis that considered an Air Tutors student to be any student who participated in at least one Air Tutors tutoring session and (b) a 24-hour dosage analysis that considered an Air Tutors student to be any student who participated in at least 24 hours of Air Tutors tutoring.

Full-Sample Analysis

The full-sample analysis compared students who participated in at least one Air Tutors tutoring session with students who did not participate in any math tutoring. This analysis provided a conservative estimate of the program's impact and reflected what happens when a program is implemented under various conditions, including partial or low participation.

Baseline Equivalence

To assess the comparability of the treatment and comparison groups in the long-term full-sample analysis, WestEd examined baseline characteristics, including student demographics and math achievement from spring 2023, which represented performance 2 years prior to the long-term outcome measure administered in spring 2025.

Table 1 presents the results of this baseline equivalence analysis for the full analytic sample (926 treatment students and 1,791 comparison students) to compare the differences between the treatment group and the comparison group.

Across all key demographic variables—including gender, economic disadvantage, English Learner status, special education status, and race/ethnicity—differences between the treatment and comparison groups were minimal. The standardized mean differences for all demographic variables were near zero, and none approached the What Works Clearinghouse threshold of 0.25 standard deviations for baseline imbalance. Taken together, these results demonstrate acceptable baseline equivalence between the treatment and comparison groups for the full-sample analysis, supporting the internal validity of the long-term intent-to-treat (ITT) impact estimates.

Table 1. Full-Sample Baseline Equivalence for Demographic Characteristics and Prior Math Achievement

Student characteristic	Treatment group (*n* = 926)	Comparison group (*n* = 1,791)	Effect size
Female	55%	53%	0.042
Hispanic	70%	70%	–0.003
White	21%	20%	0.004
Black	5%	5%	0.000
Other races	4%	4%	0.000
Economically disadvantaged	71%	71%	–0.002
English Learner	22%	22%	0.002
Special education	13%	13%	0.017
Statewide math assessment spring 2023 (standardized scale score)	–0.55	–0.52	–0.075

Source. Student records data collected from Districts A and B in the study sample.

Long-Term Impact as Measured by Statewide Standardized Summative Math Assessments

Using the full sample to estimate the long-term impact of the Air Tutors program on students' performance on statewide math assessments, WestEd compared students who attended at least one Air Tutors session during the 2023–24 school year with comparison students who did not receive any math tutoring during that same year. The analysis used student-level data from both participating districts—District A (Colorado) and District B (Texas)—and examined outcomes of the spring 2025 statewide math assessments, which were administered approximately 1 year after tutoring concluded.

WestEd first estimated the average impact of tutoring on statewide math assessment outcomes by pooling data across the two participating districts. The results show that on average, Air Tutors students scored slightly lower than did the comparison students who did not receive tutoring, but the difference was not statistically significant. The combined effect size across District A and District B was –0.035, and this difference was not statistically significant (p = .332; see Table 2 for details).

The results were then examined separately by district:

- District A: The estimated difference between treatment and comparison students was small and negative (effect size = –0.027), but this difference was not statistically significant.
- District B: The estimated difference was also small and negative (effect size = –0.043) and not statistically significant.

Summary: The long-term full-sample analysis indicates that participation in the Air Tutors program had no statistically significant impact on students' spring 2025 statewide math assessment scores. Although the direction of the estimates was slightly negative in both districts, the magnitudes were small and not distinguishable from zero. This suggests that participation in the Air Tutors program did not produce measurable long-term effects on statewide math assessment performance when all students who attended at least one tutoring session were included in the analysis.

Table 2. Full-Sample Estimates of the Impact of Air Tutors on Math Achievement Measured by Statewide Standardized Summative Assessment Scale Scores

Sample	Number of treatment students	Treatment adjusted mean (standard deviation)	Number of comparison students	Comparison adjusted mean (standard deviation)	Difference in mean	*p*-value	Effect size
Combined districts	926	–0.43 (0.63)	1,791	–0.41 (0.72)	–0.02	.332	–0.035
District A	315	–0.25 (0.68)	795	–0.23 (0.67)	–0.02	.660	–0.027
District B	611	–0.56 (0.57)	996	–0.53 (0.73)	–0.03	.363	–0.043

Note. The adjusted mean was estimated after controlling for prior math achievement (standardized score on statewide math assessment in spring 2023), grade level, gender, race/ethnicity, special education status, English Learner status, and economic disadvantage.

Source. Student records data collected from Districts A and B in the study sample.

24-Hour Dosage Analysis

The 24-hour dosage analysis focused on students who received at least 24 hours of tutoring, reflecting outcomes for students who engaged with the program at a level aligned with the dosage threshold established in the 2025 Air Tutors evaluation (Fong et al., 2025). As described in that report, the implementation plan called for students to attend 45-minute sessions 4–5 times per week for at least 10 weeks, totaling approximately 30 hours of tutoring. To meet the criterion of adequate participation (defined as 80 percent of the full program dosage), a student needed to receive at least 24 hours of tutoring.

Baseline Equivalence

To assess the comparability of the treatment and comparison groups in the 24-hour dosage analysis, WestEd examined baseline characteristics, including demographic composition and prior math achievement. Table 3 presents the results of this baseline equivalence analysis for the analytic sample consisting of 404 treatment students (those who received at least 24 hours of Air Tutors tutoring) and 1,070 comparison students. As with the full-sample analysis, the treatment and comparison groups were well matched on demographic characteristics and baseline math achievement for the 24-hour dosage analysis. These results support the baseline equivalence of the matched samples, validating that the groups were comparable on key observable characteristics prior to the tutoring.

Table 3. 24-Hour Dosage Baseline Equivalence for Demographic Characteristics and Prior Math Achievement

Student characteristic	Treatment group (*n* = 404)	Control group (*n* = 1,070)	Effect size
Female	55%	54%	.017
Hispanic	54%	55%	–.006
White	31%	31%	.007
Black	6%	6%	.000
Other races	8%	8%	.000
Economically disadvantaged	68%	68%	–.002
English Learner	15%	15%	.006
Special education	5%	5%	.000
Statewide math assessment spring 2023 (standardized scale score)	–0.42	–0.40	–.043

Source. Student records data collected from Districts A and B in the study sample.

Long-Term Impact as Measured by Statewide Standardized Summative Math Assessments

To estimate the impact of receiving at least 24 hours of Air Tutors tutoring on students' long-term math achievement, WestEd compared students who met the 24-hour threshold to comparison students who did not receive any tutoring services.

When data from the two districts were pooled, the results indicated no statistically significant differences between the treatment and comparison students. The combined effect size across District A and District B was –0.012, and this difference was not statistically significant (p = .804; see Table 4 for details).

When examined separately by district, the findings were consistent with the pooled results:

- District A: The adjusted mean for Air Tutors students was –0.24 compared with –0.23 for the comparison group, and the difference was not statistically significant.
- District B: The adjusted mean for Air Tutors students was –0.55 compared with –0.53 for the comparison group, and the difference was not statistically significant.

Summary: The 24-hour dosage analysis revealed no statistically significant long-term impacts of the Air Tutors program on spring 2025 statewide math assessment scores. These results suggest that the positive short-term effects observed on district benchmark assessments in the 2025 Air Tutors evaluation (Fong et al., 2025) did not persist through the following spring's statewide math assessments, even among students who received the recommended dosage of Air Tutors tutoring.

Table 4. 24-Hour Dosage Estimates of the Impact of Air Tutors on Math Achievement Measured by Statewide Standardized Summative Assessment Scale Scores

Sample	Number of treatment students	Treatment adjusted mean (standard deviation)	Number of comparison students	Comparison adjusted mean (standard deviation)	Difference in mean	*p*-value	Effect size
Combined districts	404	–0.32 (0.67)	1,070	–0.31 (0.76)	–0.01	.804	–0.012
District A	298	–0.24 (0.68)	769	–0.23 (0.79)	–0.01	.887	–0.008
District B	106	–0.55 (0.60)	301	–0.53 (0.66)	–0.02	.767	–0.028

Note. The adjusted mean was estimated after controlling for prior math achievement (standardized score on statewide math assessment in spring 2023), grade level, gender, race/ethnicity, special education status, English Learner status, and economic disadvantage.

Source. Student records data collected from Districts A and B in the study sample.

Discussion

The long-term findings presented in this brief provide more context for understanding the persistence of the Air Tutors program's effects on student math achievement. In the 2025 Air Tutors evaluation (Fong et al., 2025), it was found that the short-term impacts of the Air Tutors program varied across the outcome measure and analytic samples:

- In the full-sample analysis, short-term impacts on statewide standardized summative math assessments were negative and statistically significant for the combined sample (both Districts A and B) and for District B only. By contrast, impacts on district benchmark assessments were positive and statistically significant in District A and positive but not statistically significant in District B.
- For the 24-hour dosage analysis, there were no statistically significant impacts on statewide math assessments, but there were positive and statistically significant gains on district benchmark assessments in both districts, highlighting the program's potential to improve math achievement when students received sufficient tutoring dosage.

The follow-up analysis provided in this brief extends that work by examining whether those impacts were sustained 1 year later on the spring 2025 statewide math assessments.

Across both the full sample and the 24-hour dosage sample, the long-term results show no statistically significant differences between the treatment students who participated in the Air Tutors program and the comparison students who did not receive tutoring. Effect sizes were small in magnitude and not statistically distinguishable from zero, both in the combined sample and when examined separately by district. These findings indicate that, although the program provided short-term support for math learning, the effects did not persist through the subsequent school year as measured by statewide standardized summative tests.

Several factors may help explain the results. First, the timing of the tutoring relative to the long-term assessment (i.e., spring 2025 statewide math assessments) is an important consideration. The Air Tutors sessions in both districts delivered during the 2023–24 school year concluded approximately 1 year before the 2025 state testing window. Prior research on high-dosage tutoring (e.g., Robinson & Loeb, 2021; Max & Place, n.d.) underscores the importance of sustained, frequent instructional contact for producing and maintaining learning gains. These studies and field reports suggest that when tutoring concludes well before end-of-year assessments or is delivered inconsistently over time, students may have fewer opportunities to reinforce or retain the skills developed during tutoring. The absence of detectable long-term

effects in this evaluation is therefore consistent with expectations from the tutoring literature regarding the challenges of sustaining short-term gains once intensive support ends. It is also noteworthy that both districts experienced substantial leadership and staffing turnover during and after the implementation period and reductions in professional development supports, which may have shaped the broader instructional context during the follow-up year as well.

Second, dosage and implementation variability likely reduced cumulative impact. As discussed in the 2025 Air Tutors evaluation (Fong et al., 2025), student participation in District B remained far below the intended 30 hours because scheduling disruptions and session cancellations limited consistent exposure. Although the 24-hour threshold analysis isolated students who met recommended participation levels, dosage patterns across the broader sample suggest that many students did not receive sustained tutoring, which limited the potential for lasting gains beyond the immediate instructional period.

Third, alignment between tutoring content and subsequent assessments may have constrained long-term effects. Because the tutoring sessions were designed to support the 2023–24 school year's grade-level standards, the material covered may not have directly mapped to the math content assessed 1 year later. This lack of vertical alignment between intervention content and later state assessments could partially explain an absence of sustained gains.

Finally, differences in assessment focus may partially account for the divergence between short- and long-term outcomes. Adaptive benchmark assessments are designed to capture immediate skill acquisition and incremental growth within the tutoring period, whereas statewide standardized summative assessments emphasize cumulative mastery across the full curriculum. As noted in one systematic review and meta-analysis of the experimental evidence of tutoring on preK–12 learning (Nickow et al., 2020), tutoring effects are typically measured immediately after the intervention and tend to be most evident on assessments that are closely aligned in both timing and instructional content with the tutoring itself. Therefore, the absence of long-term gains on statewide math assessments does not necessarily contradict the earlier benchmark improvements but, rather, underscores the challenge of sustaining short-term learning gains over time.

In summary, the results from this long-term follow-up study indicate that the Air Tutors program produced no measurable impact on statewide standardized summative math performance 1 year after tutoring. However, given prior evidence of short-term gains and the importance of dosage and fidelity in high-dosage tutoring models, these findings point to the need for sustained, well-timed implementation rather than to the ineffectiveness of the tutoring model itself. Future efforts should focus on ensuring continuous participation, extending tutoring so that it is closer to the testing window, and reinforcing alignment with classroom instruction to translate early gains into enduring academic progress.

References

Fong, A., Chen-Gaddini, M., & Huang, M. (2025). *An evaluation of the Air Tutors program: Findings from an Arnold Ventures grant*. WestEd. https://www.wested.org/resource/an-evaluation-of-the-air-tutors-program/

Max, J., & Place, K. (n.d.). *Accelerate's first call to effective action: A synthesis of lessons learned.* Accelerate. https://accelerate.us/wp-content/uploads/2023/12/Lessons-Learned-from-Accelerate-CEA-2022-23-1.pdfhttps://accelerate.us/wp-content/uploads/2023/12/Lessons-Learned-from-Accelerate-CEA-2022-23.pdf

Nickow, A. J., Oreopoulos, P., & Quan, V. (2020). *The impressive effects of tutoring on preK–12 learning: A systematic review and meta-analysis of the experimental evidence* (EdWorkingPaper No. 20-267). Annenberg Institute at Brown University. https://doi.org/10.26300/eh0c-pc52

Robinson, C. D., & Loeb, S. (2021). *High-impact tutoring: State of the research and priorities for future learning* (EdWorkingPaper No. 21-384). Annenberg Institute at Brown University. https://doi.org/10.26300/wghb-4864