2024-2025 ANALYSIS OF HIGH DOSE TUTORING PROGRAM IN BIRMINGHAM CITY SCHOOLS

The implementation and impact of college students as school-based tutors

Abstract

High Dose Tutoring has been shown by research to improve student outcomes. Birmingham City Schools is continuing to improve implementation and data-tracking of it tutoring efforts. While some promising signs can be seen, the effort to measure the impact of the program is complicated by a small sample size. Comparing HDT Students to students in the same grade that did not receive HDT, HDT made greater score gains in math in grades 6-8. For students tutored in ELA, results were mixed. Evaluation of impact on high school students is not available for the lack of a measurement tool in those grades.

Thomas Spencer & Jason Fulmore
thomas@parcalabama.org

High Dose Tutoring Evaluation

Overview: Due to small sample sizes and other complications in building comparisons, PARCA's analysis of the impact of High-Dose Tutoring cannot measure with confidence the impact on test scores produced by high-dose tutoring. However, the overall 2024205 results of a simplistic comparison of HDT students and the rest of the students in the same grade are consistent with those from the previous year. In Math, in grades 6-8, students receiving high-dose tutoring made higher average score gains than students who did not. In ELA, the results were mixed. This year, in addition to test score results, PARCA added an analysis of payroll records to provide program administrators with insight into trends in the number of tutors, the number of hours of tutoring billed, and the compensation paid to tutors.

The Birmingham School System has sponsored a program that provides tutoring for students, primarily in middle and high school, with college students serving as paid tutors for the sessions. The aim is to improve student performance by pairing students with knowledgeable students who are close to their own age for academic enhancement and inspiration. Secondly, it is hoped that participating college students will be attracted to the teaching profession and to Birmingham City Schools as a result of their involvement.

In national education research, high-dose tutoring has been shown to improve student outcomes. The Center for American Progress[1] defines High Dose Tutoring as:

- **One-on-one or small-group sessions with no more than four students per tutor**
- **Use of high-quality materials that align with classroom content**
- **Three tutoring sessions per week—at minimum—each lasting at least 30 minutes**
- **Sessions held during school hours**
- **Students meeting with the same tutor each session**
- **Professionally trained tutors who receive ongoing support and coaching**

Birmingham City Schools continues to work toward these aspirations; however, the program, as currently constituted, does not completely conform to the description outlined in the literature. However, the program continues to work toward an implementation that reflects aspects of the high dose definition.

The 2024-2025 Evaluation of High-Dose Tutoring in Birmingham City Schools finds that data collection continues to improve. More sessions of High-Dose tutoring were recorded (2,419 in 2024-2025 vs. 1,193 in 2023-2024). Those logged sessions also included more complete information about the content of the tutoring session.

The collection of student ID numbers of participating students also improved. The number and percentage of students we were able to identify as having participated climbed. However,

[1] ("Scaling Up High-Dosage Tutoring Is Crucial to Students' Academic Success" 2024)

some schools still do not identify participating students with a unique student ID number, making matching difficult.

At the same time, challenges persisted in assessing the impact of HDT. While the number of students identified by SSID increased, it remained incomplete, which reduced the sample size.

The sample was also decreased because some students did not have a baseline and an end-of-the-year iReady score. Growth on iReady is the most obvious way to chart academic growth.

We also lack a standardized measure for tracking growth among students in grades 9 through 12. It is possible that we can use pre-ACT and ACT scores across years to establish a before-and-after measure. However, that was not available for this analysis.

In middle school grades, using iReady scores, we attempted to build a comparison set of students who were at the same school in the same grade and who started with a similar baseline score as the HDT students. This proved to be impossible in some instances due to the small sample size and because, at some schools and in most cases, the students participating in HDT had high baseline scores.

Having higher-scoring students dominating participation in HDT presents two problems. One issue is that it becomes difficult to construct a "non-treatment" group of students with similar characteristics. Secondly, high-scoring students in general are not as likely to make large score gains. Students who start with a lower baseline score have more room to grow. A detailed analysis of results by grade is included after the summary results.

Summary Results

To the extent that we could build comparison datasets, HDT students made higher gains in math, while ELA score gain comparisons were mixed. However, none of the results produced statistically significant results that would allow us to say that HDT led to disproportionate score gains.

We present the following results with the understanding that educational interventions do not occur in a laboratory, and thus, the conditions for measurement aren't perfect, and our ability to draw firm conclusions is limited.

Figure 1 compares the iReady results of students who received high-dose treatment and those who did not. The matched comparison group are students from the same school who had similar baseline scores as the HDT students. The unmatched group consists of all the students in the system who are not in the matched group for that grade. In all three grades, the tutored students made greater gains than the comparison students and the unmatched students.

Examining the baseline scores of the students in the three grades reveals that each grade varies. All three groups have nearly the same average baseline score in 6th grade. In 7th grade, HDT students start behind, but gain more than the comparison groups. In 8th grade, HDT students start higher than the average student, but lower than the comparison group; however, in the end, the HDT students make much larger gains. It should be noted that the HDT and comparison group are much smaller in 8th grade than in the other grades. See Figure 2.

FIGURE 1. MATH GAINS, BASELINE AND POST SCORES

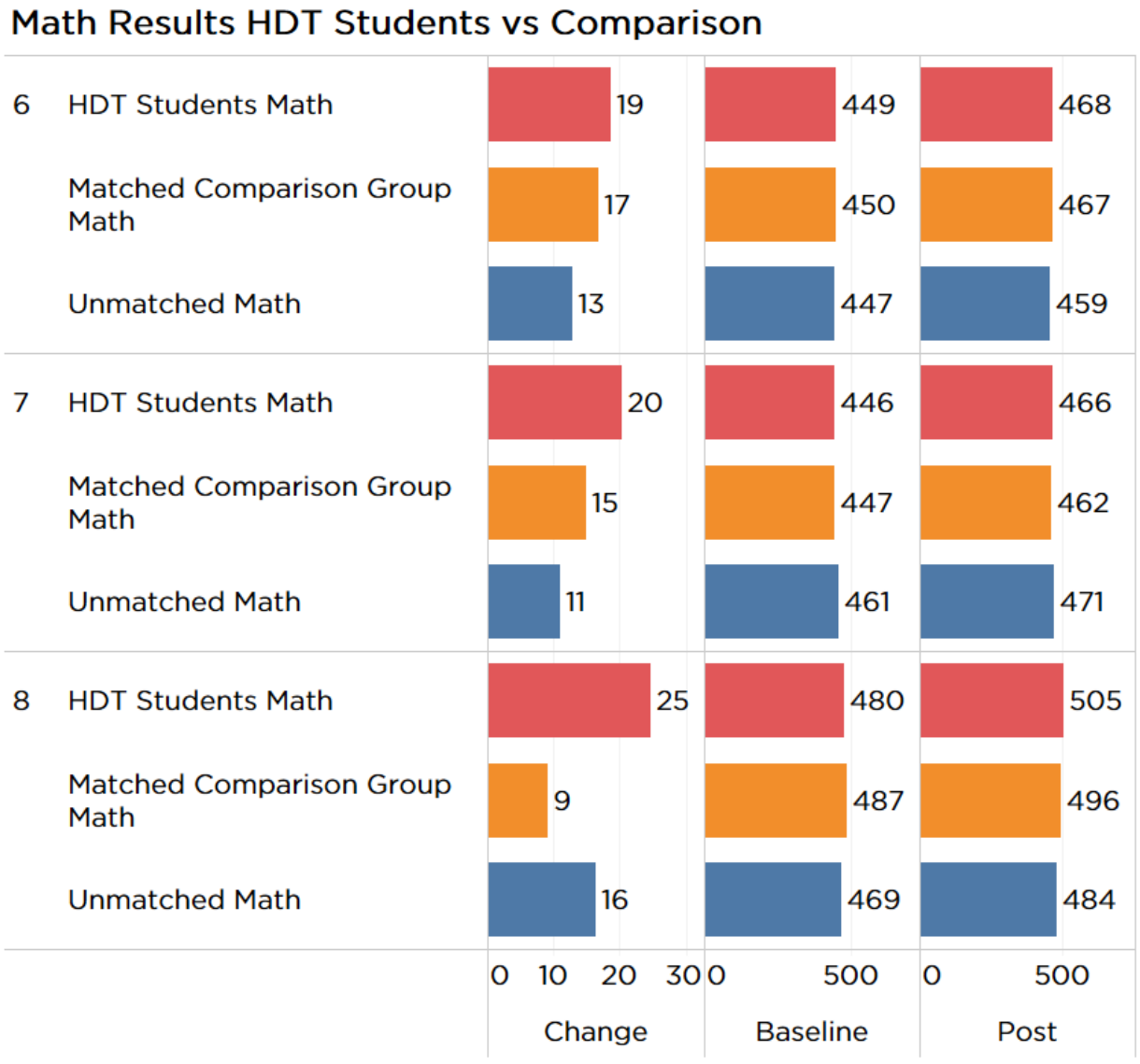
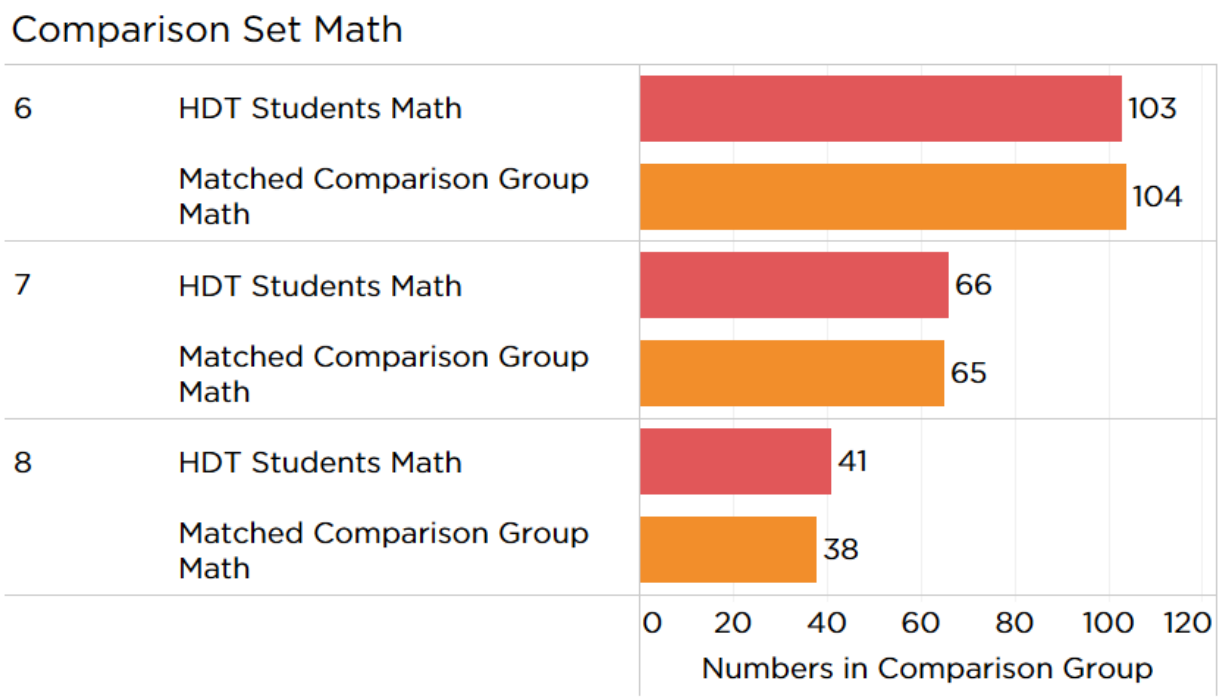


FIGURE 2. NUMBERS IN THE COMPARISON GROUP



English Language Arts

ELA results were mixed. The sample sizes were smaller. In 8th grade, the sample was too small to form a comparison group. Thus, the same cautions apply: while we can share observations, we are unable to draw conclusions.

The HDT students in both 6th and 7th grade start at a higher baseline on average. In 6th grade, the HDT students gained less than the comparison group and the unmatched students.

In 7th grade, the HDT students outgained the matched comparison group. The HDT group had an average baseline that was significantly higher than that of the rest of the students in the grade. The unmatched students gained more, but HDT students ended the year with a much higher finishing score.

In 8th Grade, only 10 students participated in ELA HDT and had both pre- and post-iReady scores. The participants in HDT had lower average baseline scores but achieved a significantly higher level of gain than the average 8th-grade student on the ELA iReady.

FIGURE 3. ELA GAINS, BASELINE AND POST SCORES

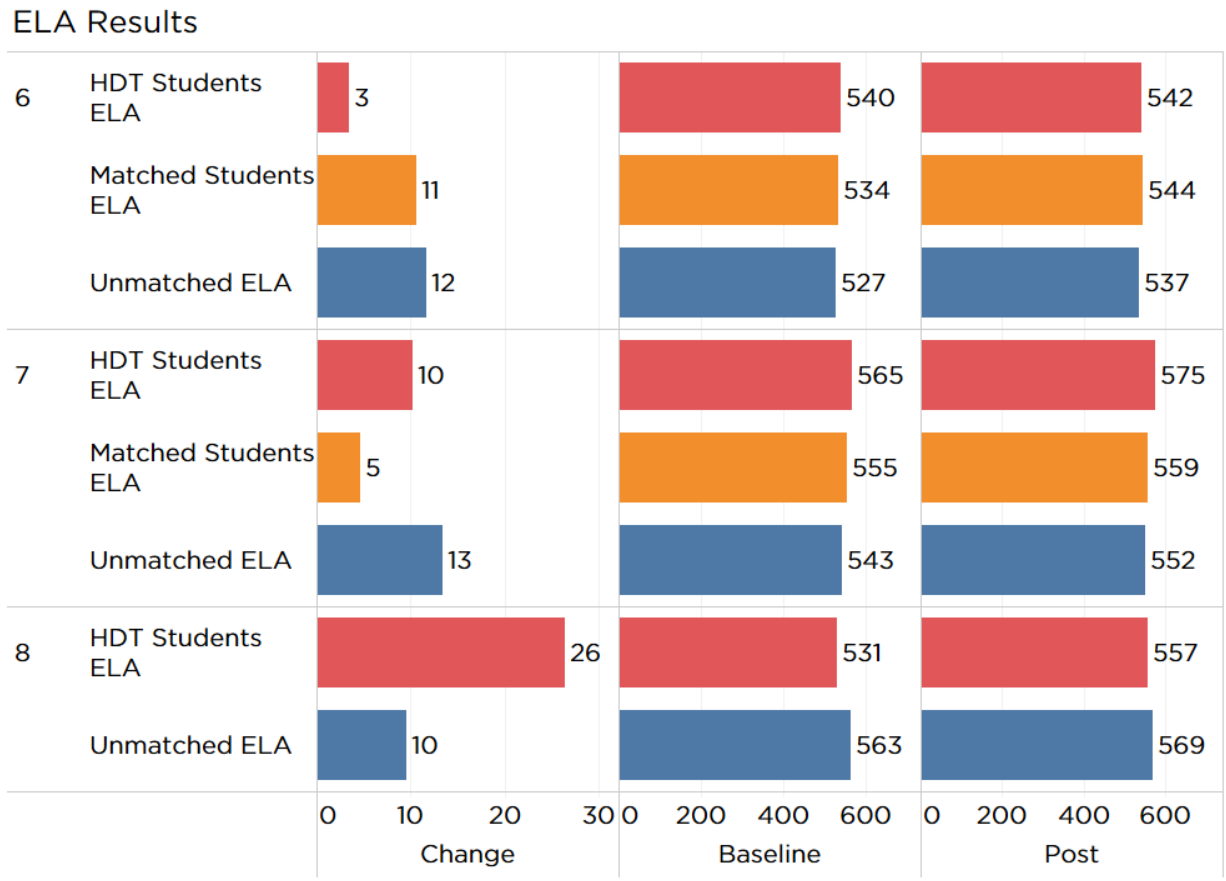
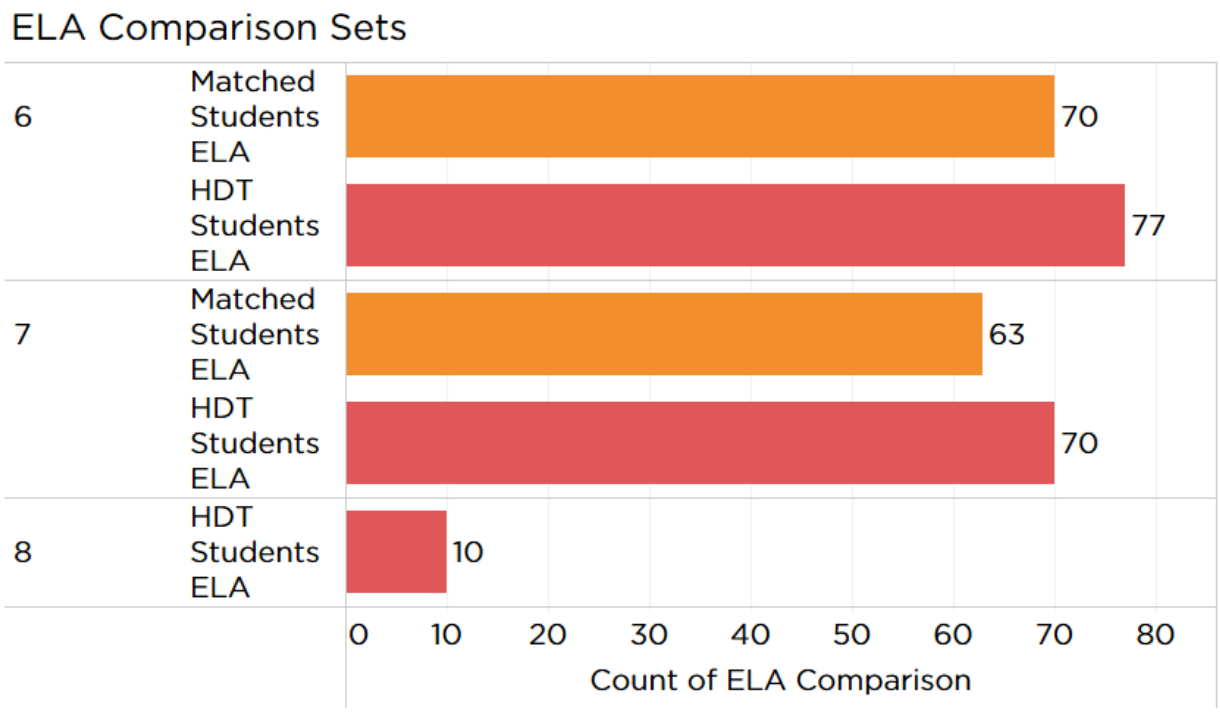


FIGURE 4. NUMBERS IN THE COMPARISON GROUP



It is recommended that Birmingham City Schools' high-dose tutoring continue to improve the system for gathering information on participating students.

We recommend identifying growth metrics for high school students so that HDT students in high school can be evaluated for their impact.

We do not recommend exclusive targeting of either high-scoring or low-scoring students for HDT participation. The choice of targeting is up to the district and/or the school and should be grounded in strategic choices that may be unique to each school.

Trends in Compensation, Hours, and Tutor Recruitment

New to the analysis this year is an analysis of payroll records for the program. The records provide insight into the number of tutors, the hours they billed, and their total and average compensation.

Of the three years examined, each year has a different pattern. More tutors participated in 2023, but their median hours worked were lower than in 2024 and equal to 2025. Fewer tutors participated in 2024, but of those who did, the average number of hours spent tutoring was much higher. Both in 2023 and 2024, a handful of tutors logged an exceptional number of hours. Some of the same tutors who logged over 200 hours of tutoring in 2024 had billed for a similar number of hours in 2023. In 2025, the number of tutors increased over 2024, but their average hours billed declined.

Program administrators should combine this information with a frontline understanding of the differing approaches to recruiting and deploying tutors. Other factors to consider in the evaluation include how payroll records and tutor management affected the totals.

FIGURE 5. NUMBER OF TUTORS BY YEAR

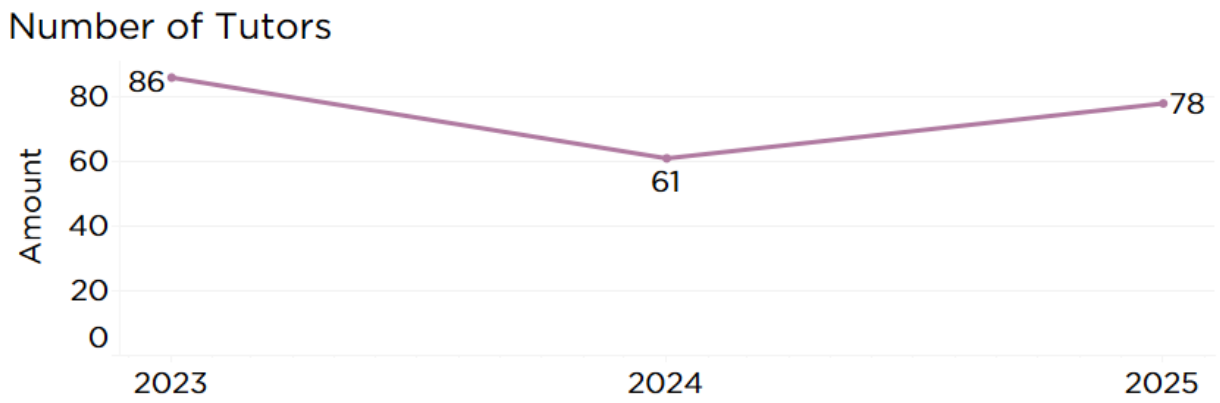
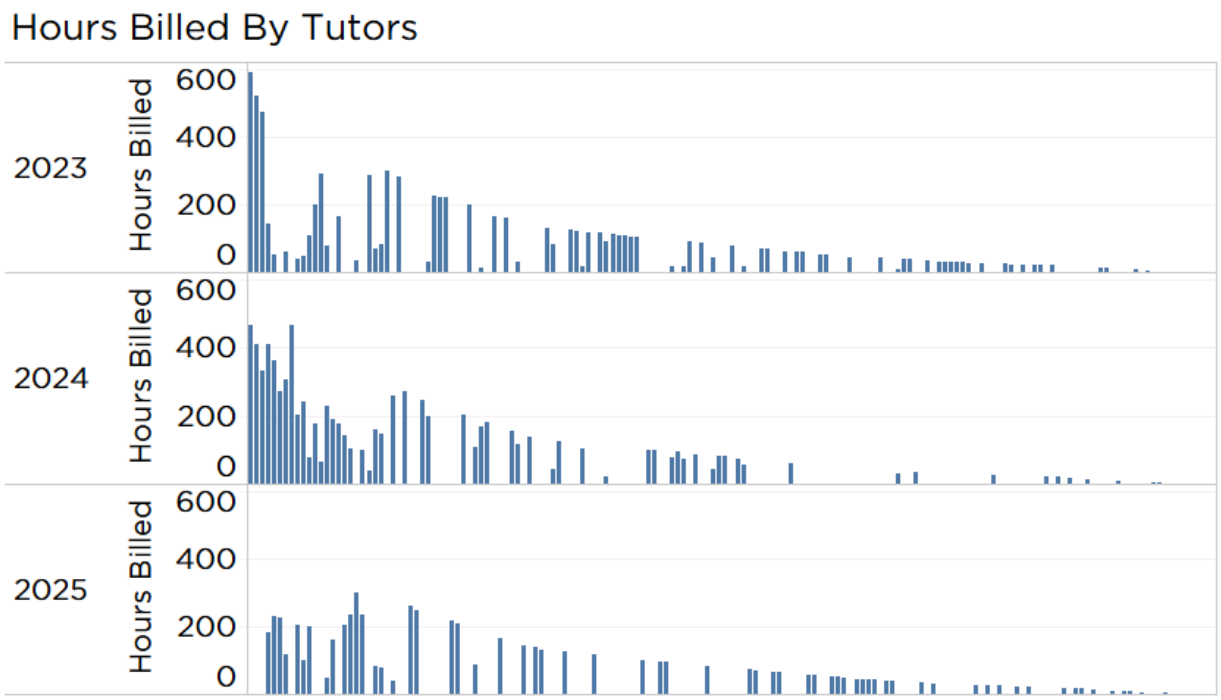


FIGURE 6. HOURS BILLED BY TUTORS BY YEAR



Each line represents an individual tutor. The position of those individual lines is constant over multiple years if the tutor taught in multiple years. Notably, a cluster of the same tutors from 2023 and 2024 delivered a high number of hours.

Figures 7 and 8 present hours and compensation for tutors over the past three school years. Billed tutoring hours peaked in 2024, despite there being fewer individual tutors than in 2023 or 2025.

FIGURE 7. TOTAL, AVERAGE, AND MEDIAN TUTOR HOURS

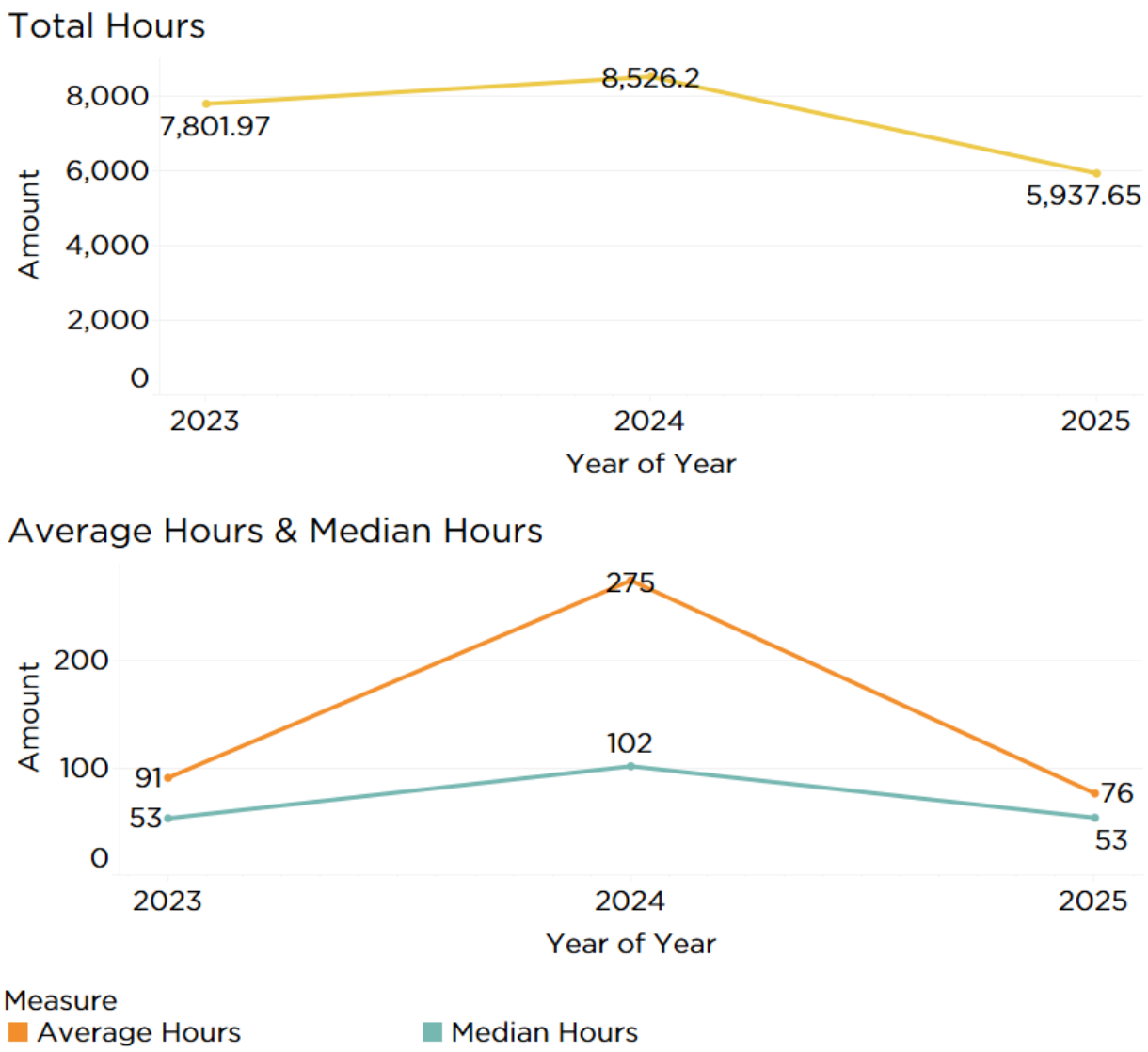
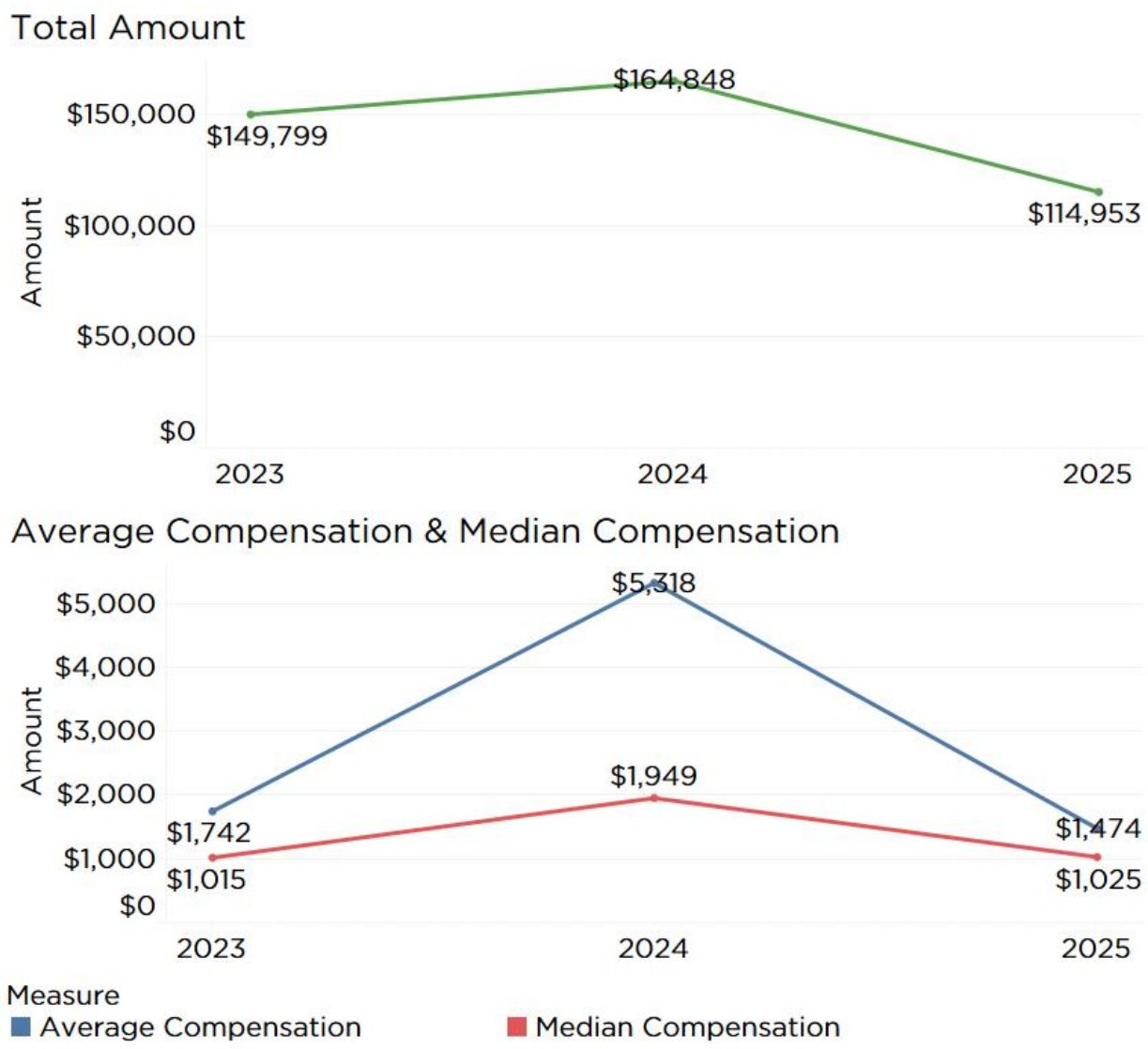


FIGURE 8. TOTAL AND AVERAGE TUTOR COMPENSATION



Detailed Results based on iReady Results

HDT Analysis 2024-2025 using iReady Data

Sample and Data Provided by Birmingham City Schools

iReady data was gathered at three times during the 2024-2025 school year. For this analysis, we will focus on changes over time at timepoints (baseline and post). Baseline is defined as Baseline Diagnostic (Y/N) with a Y AND Completion Date prior to 10/31 of the academic year, and post-test is defined as having a Completion Date between 3/2 and 7/1 of the academic year. The post-test is also the "Most Recent Diagnostic YTD (Y/N)" and will not have a Y for Baseline Diagnostic (Y/N).

A Google sheet "HDT Student List" was provided to PARCA on June 18, 2025, by Constance Blaylock of Birmingham City Schools that contained 16 worksheets (one for each school) and a total of 525 students. The students received tutoring for different things … that breakdown is below:

Content Area	N Students
ACT	9
ACT Prep	19
ELA	220
Math	241
Workkeys	36
Total	525

The iReady assessment focuses on ELA and Math and is primarily used by elementary and middle school students. Only a small number of high school students complete the iReady assessment as compared to elementary and middle grade students. For this analysis, we will focus on ELA and Math (n=461), but keep in mind that the total sample will be smaller than 461 because some high school students participated in ELA and Math tutoring (n=24), and the likelihood of having iReady scores for those students is small.

A more comprehensive appraisal of High Dose Tutoring's impact would be possible if we had pre- and post-assessment measures for high school students.

Two additional notes regarding the construction of the dataset. First, a SSID (Student ID) was not provided for all students. The SSID is the variable used to match students to the iReady data. Without it, we may not have been able to match students who participated (lowering our overall sample size).

Second, several students participated in more than one type of tutoring. If a student participated in both ELA and Math, then they are included in both analyses.

The HDT Student List dataset has columns for January, February, March, April, and May, but no data in those fields except for February. As a result, we have dichotomized the data into 0 = a matched participant (non-HDT student) and 1 = a student who participated in High-Dose Tutoring, and we are making the assumption that their name on the list indicates they attended at least some

tutoring sessions. The other measures provided were grade (in some cases), school, and the tutor's name. The analysis will be conducted by grade level due to the limitations of the assessment, but may or may not include an analysis by school or tutor, depending on the small sample size.

Construction of the HDT2425 ELA Dataset

The HDT Student List was combined with the iReady dataset (ELA 2425 Simple Matched and Math 2425 Simple Matched). Students were identified and provided a 1 in the HDT variable to indicate them as participants. They were matched to another student based on three variables: grade, school, and baseline iReady score (within half a standard deviation of the group). The purpose of matching to a single student is to create a comparison group of students who are similar to those who participated in High Dose Tutoring.

The original data provided had 220 students who participated in ELA tutoring. When combined with ELA2425 Simple Matched (iReady data for 2425), we lose 60 students who did not have baseline iReady scores. The other 160 are spread across three grades – 77 in 6th grade, 73 in 7th grade, and 10 in 8th grade, as well as nine schools. After adding those who completed a post-iReady assessment, our sample decreases to 75 students in 6th grade across five schools, with only Huffman Middle School having a large enough sample to assess change within the school. At the 7th-grade level, the sample size decreases from 73 to 67 when including students who have taken the iReady post-assessment, with Wilkerson Middle being the only school with a large enough sample to examine changes at the school level.

ELA Analysis – 6th Grade (Overall)

N= 75 for HDT; N=70 for Comparison

Statistically, there was no difference between the HDT group and a matched comparison group of 6th-grade students.

Both groups showed a statistically significant change from baseline to the post-assessment, but no difference between the groups.

One challenge of this analysis was that one school (Inglenook) had 7 of their top 10 6th graders (on the baseline iReady) participate in high-dose tutoring, and we did not have any 6th graders in Inglenook to match them with. Removing them from the analysis would not have likely changed whether there was a difference between the two groups. Still, it is noted to illustrate one challenge of the data (some higher performing students are the students seeking out the tutoring).

75 HDT – Baseline was 540.31 and Post was 543.80 – no statistical change per a paired samples t test

70 Comparison- Baseline was 533.50 and Post was 544.16 – statistically significant change over time

All 6th – Baseline was 528.44 and Post was 539.56 (n=1305)

ELA Analysis – 6th Grade (Huffman)

N= 29 for HDT; N=26 for Comparison

There is no statistical difference between the two groups, and no time effect is observed for either group. As noted above, Huffman Middle School is another example of HDT attracting the highest performing students (at least in terms of Baseline iReady scores).

The HDT group mean at Baseline was 551.55, compared to the Comparison group mean of 541.81. Once again, 7 of the top 8 students (as measured by the Baseline iReady) participate in HDT. Finding an appropriate comparison group of students is a challenge again, this time at the school level.

ELA Analysis – 7th Grade (Overall)

N= 63 for HDT; N=67 for Comparison

No statistical difference was found between groups, but a time effect was observed for both groups combined. We have the sample issues in the 7th grade, as we saw in the 6th grade. Overall, the HDT group had a baseline score that was 11 points higher than the comparison group (565.31 vs. 554.70). The comparison group gained approximately 5 points from baseline to post. In contrast, the HDT group gained 10 points (the difference is not statistically significant using a Repeated Measures ANOVA with a $p \leq .05$).

HDT – 565.31 to 575.57 – statistically significant

Comparison – 554.70 to 559.37 – not statistically significant

Overall – 544.46 to 557.26 (n=1201)

ELA Analysis – 7th Grade (Wilkerson)

The 7th-grade students at Wilkerson are the most illustrative example of the sample and matching issues we have seen thus far. 23 of the top 30 7th graders at Wilkerson (as measured by their baseline iReady score) participated in high-dose tutoring. None of the bottom 30 students participated in high-dose tutoring. Making a comparison between the participating and non-participating groups is problematic (as we did above for all 7th graders). Still, as long as we continue to have top students participating and bottom students not participating, we will continue to be challenged in assessing the impact of high-dose tutoring.

Construction of the HDT2425 Math Dataset

We used the same process for matching students between the HDT dataset and the Math2425 Simple Matched dataset. We began with 237 students, but lost 22 because they did not have a baseline assessment in place. That leaves us with 215 students over three grades (6th, 7th, and 8th) from 10 different schools.

Frequency count:

School	6th	7th	8th	Total
Booker T Washington K8	26	0	0	26
Bush Hills Steam Academy	15	16	12	43
Hayes K8	0	15	0	15
Inglenook K8	9	8	12	29
Phillips Academy	6	0	7	13
Ossie Ware Mitchell Middle	0	16	13	29
Smith Middle	18	0	0	18
South Hampton	22	0	0	22
WJ Christian K8	7	0	0	7
Wilkerson Middle	1	12	0	13
Total	104	67	44	215

Math Analysis – 6th Grade (Overall)

N= 100 for HDT; N=104 for Comparison

Statistically, there was no difference between the HDT group and a matched comparison group of 6th-grade students.

Both groups showed a statistically significant change from baseline to the post-assessment, but no difference between the groups. Unlike the challenge in ELA, the two groups have similar baseline scores (449.82 vs. 449.39), and their change was nearly identical, with the HDT group gaining approximately 18 points and the comparison group gaining approximately 16 points.

The only school with a large enough sample in 6th Grade is Booker T Washington, and a Repeated Measures ANOVA shows no difference between the groups as well (same as the overall analysis).

HDT – Baseline – 449.39 – Post – 468.04 (Statistically significant)

Comparison – Baseline – 449.82 – Post – 466.55 (significant)

Overall – Baseline – 447.81 – Post – 461.34 (n=1308)

Math Analysis – 7th Grade (Overall)

N= 63 for HDT; N=65 for Comparison

Statistically, there was no difference between the HDT group and a matched comparison group of 7th-grade students.

Both groups showed a statistically significant change from baseline and the post assessment, but (statistically) no difference between the groups. The HDT group showed more growth from baseline to post (20.28 vs. 14.95), but statistically they are not different at $p \leq .05$.

HDT – Baseline – 445.59 – Post – 465.87 (significant)

Comparison – Baseline – 447.40 – Post – 462.35 (significant)

Overall – Baseline – 459.76 – Post – 471.48 (n=1206)

Math Analysis – 8th Grade (Overall)

N= 40 for HDT; N=38 for Comparison

Statistically, there was no difference between the HDT group and a matched comparison group of 8th-grade students.

Both groups showed a statistically significant change from baseline and the post assessment, but (statistically) no difference between the groups. The HDT group showed more growth from baseline to post (24.65 vs. 9.05), but statistically, the difference was not significant at $p < .05$. The small sample size likely played a role in whether the two groups were statistically different in 8th grade.

HDT – Baseline 479.93 – Post – 504.58

Comparison – Baseline 486.82 – Post – 495.87

Overall – Baseline 470.29 – Post 486.77