**Summary of Findings on the Impact of a Data Science high school course on student mathematics enrolment, mathematics achievement, and excitement for STEM .**

**June 30, 2023**

**Introduction**

This report details the findings of a year-long study of students taking data science in 5 high schools across multiple districts in California. The research questions are:

* + Do student learning outcomes differ between Data Science students and Algebra 2 students, in an assessment of content taught in both courses?
  + Do more students enroll in 3rd and 4th year mathematics courses once the high school data science course is offered?
  + Are course-taking patterns in 3rd and 4th year mathematics more diverse in student composition (comparing data science to other advanced math courses)?
  + Do students who take the high school data science course see a future for themselves in STEM?
  + Do students who take the high school data science course report that the learning is more meaningful and relevant than in previous mathematics courses?

The study found that students who took the high school data science course:

* Scored at significantly higher levels on a test of algebra with data, than students in algebra 2 classes.
* Enrolled in significantly more math courses than comparable students not taking data science.
* Were significantly more likely than peers who had not taken the course to see a future for themselves in STEM, have an interest in STEM, and have confidence in their data science skills relevant to STEM.
* Reported that their learning in the data science course was more meaningful and relevant than in previous mathematics courses.

Table 1 provides a summary of demographic variables for each of the research sites.

Table 1. *Demographic Profiles of Participating Schools\**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Total School Enrollment | Percent  Race/Ethnicity | Percent  Low SES | Percent at or above Math Proficiency (State exam, 2022) |
| HS 1 | 644 | 14.9 White  80.2 Latinx  2.2 Black | 87.9 | 13.68 |
| HS 2 | 1,568 | 15.4 White  79.3 Latinx  1.5 Black | 63.1 | 12.91 |
| HS 3 | 602 | 12.0 White  82.9 Latinx  0.8 B Black | 77.2 | 41.03 |
| HS 4 | 2,914 | 54.4 White  22.8 Latinx  1.0 Black | 31.2 | 75.16 |
| HS 5 | 1,549 | 28 White  56.7 Latinx  1.9 Black | 50.9 | 23 |

*\* Data retrieved from California State Department of Education*

The course was implemented in five high schools, with a total of six teachers teaching one section of Data Science during the 2021-22 school year.

The study found that data science students were more racially and socioeconomically diverse and had slightly lower prior achievement in algebra than students enrolled in the other higher level math classes. Taken together, these results suggest that a full year data science course offered in the 3rd and 4th year of high school is a powerful means of encouraging a larger and more diverse group of students to engage and persist in mathematics and STEM. Given the needs of the 21st century workforce and the urgency of updating mathematics pathways in US high schools, these findings indicate that a full year data science is a promising approach to address these challenges.

1. **What Are Students Learning in Data Science, compared with Algebra II?**

Students taking Algebra II or Data Science were given an assessment that combined concepts in Algebra and Data Science. The first question is from NAEP and includes mathematics standards from Algebra. 2 The second question is an assessment of linear functions, content that appears in Algebra I and II. The assessment was administered to approximately 480 students.

Assessment Item 1

|  |
| --- |
| The question requires you to show your work and explain your reasoning. You may use drawings, words, and numbers in your explanation. Your answer should be clear enough so that another person could read it and understand your thinking. It is important that you show all of your work.  The table below shows the number of people each week getting a vaccination in 2 regions. The mean (average) and the median values are shared at the end of the table.  A screenshot of a table  Description automatically generated with low confidence  (a) Which statistic, the mean or the median, would you use to describe the vaccination numbers for the 5 weeks in region A? Justify your answer.  (b) Which statistic, the mean or the median, would you use to describe the vaccination numbers for the 5 weeks in region B? Justify your answer. |

The second question is adapted from the textbook: Core Plus Mathematics Course 1, Unit 3 Linear Functions.

Assessment Item 2

|  |
| --- |
| The data in the table linked below shows how average daily food supply (in calories) is related to life expectancy (in years) and infant mortality rates (in deaths per 1,000 births) in a sample of countries.  A picture containing text, screenshot, font, number  Description automatically generated  (a) Make scatter plots of the (Daily calories, life expectancy) and (daily calories, infant mortality) data. Feel free to use any tools you would like to create these plots. Study the patterns in the table and scatter plots use them to answer the following questions.  (b) If you used google sheets please share a link with permission to view your plot. To change the share settings, click on the share button in the top right corner and set to "anyone with link can view."  (c) What seems to be the general relationship between daily calories and life expectancy in the sample countries?  (d) Economists might use a linear model to predict the increase of life expectancy or decrease of infant mortality for various increases in food supply. Determine a linear model for calculating life expectancy from calories using the (daily calories, life expectancy) daily pattern.  (e) What story does this data tell?  (f) What other data would you like to collect to draw stronger conclusions? |

The total score used for the main analysis is calculated as the sum of the scores on the first question and each of the sub-questions of the second question. The Cronbach’s Alpha reliability of the total score is .74 suggesting the assessment is a relatively consistent measure of student ability.

Results

Analysis of Covariance (ANCOVA) is used to consider how students taking Data Science compared to students taking Algebra II on an assessment of algebra and functions, with data. A series of covariates are used to control for pre-existing differences between the two groups. These include most recent prior math achievement (whichever of Algebra 1 or Algebra 2 Grade Point Average (GPA) that was available, grade level, age, attendance, sex, indicators of race and ethnicity, Free and Reduced Lunch (FRL) status, English Language Learner (ELL) status, special education status, and whether students had ever been suspended. Table 2 shows the means of the covariate variables for the students enrolled in Algebra and Data Science courses available for the analysis.

Table 2. *Descriptive Statistics by Course*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Algebra | | Data Science | | | Total | | |  |
|  | Mean | Std. dev. | | Mean | Std. dev. | | Mean | Std. dev. | |
| Most Recent Math Achievement (n = 478) | 2.85 | .97 | | 2.52 | 1.10 | | 2.80 | 1.00 | |
| Grade Level (n = 474) | 10.69 | .58 | | 11.62 | .49 | | 10.84 | .66 | |
| Age\* (n = 481) | 16.89 | .65 | | 17.80 | .65 | | 17.03 | .73 | |
| Attendance (n = 474) | .95 | .06 | | .94 | .05 | | .95 | .06 | |
| Female (n = 480) | .57 | .50 | | .46 | .50 | | .55 | .50 | |
| White (n = 481) | .73 | .45 | | .76 | .43 | | .73 | .44 | |
| Black (n = 481) | .02 | .16 | | .00 | .00 | | .02 | .14 | |
| American Indian or Alaska Native (n = 481) | .06 | .25 | | .13 | .34 | | .07 | .26 | |
| Asian (n= 481) | .11 | .31 | | .07 | .25 | | .10 | .30 | |
| Pacific Islander (n = 481) | .01 | .09 | | .00 | .00 | | .01 | .08 | |
| Other race (n = 481) | .07 | .25 | | .04 | .20 | | .06 | .25 | |
| Hispanic/Latinx (n = 481) | .29 | .45 | | .43 | .50 | | .31 | .46 | |
| Free or Reduced Lunch (n = 474) | .31 | .46 | | .43 | .50 | | .33 | .47 | |
| English Language Learner (n = 474) | .02 | .13 | | .08 | .27 | | .03 | .16 | |
| Special Education Student (n = 474) | .03 | .17 | | .00 | .00 | | .03 | .16 | |
| Suspensions (n = 474) | .00 | .05 | | .00 | .00 | | .00 | .05 | |

\* Age is calculated as of May 31, 2022

Based on independent sample t-tests, Data Science students had lower prior math grades [*t* (476) = -2.67, *p* < .01], were in a higher grade [*t* (472) = 13.08, *p* < .05], were older [*t* (479) = 11.27, *p* < .05], more likely to be American Indian or Alaska Natives [*t* (478) = 2.05, *p* < .05], were more likely to be Hispanic [*t* (472) = 2.52, *p* < .05], were more likely to be enrolled in an FRL program [*t* (472) = 2.12, *p* < .05], and were more likely to be English Language Learners [*t* (472) = 3.02, *p* < .01].

After removing outliers, a total of 427 students (71 in data science and 356 in Algebra II) who completed the assessment and for whom the districts were able to provide all the required demographic data were used for the main analysis. The ANCOVA analysis revealed Data science students had a higher covariate adjusted mean total score than the Algebra students, *M*Data Science = 19.02, (SE = .64) and *M*Algebra = 15.98 (SE = .52)[[1]](#footnote-1). This effect is significant, *F*(1, 408) = 13.63, *p* <.001. The data for this analysis met the necessary assumption of equality of variance. Figure 1 shows the estimated covariate adjusted total score means and their 95% confidence intervals.

Figure 1. *Covariate adjusted total score means for Data Science and Algebra students*A blue and green rectangles

Description automatically generated with low confidence

Covariates appearing in the model are evaluated at the following values: most recent Alg = 2.8482, sex = .56, grade2021\_22 = 10.83, Age = 17.0233, FRL21\_22 = .32, ELL21\_22 = .02, SPED21\_22 = .03, Attend21\_22 = .9495, Susp21\_22 = .00, Black = .02, Amer Ind = .08, Asian = .11, Native Hawaiian = .01, Other Race = .06, Ethnicity = .31.

1. **Number of Mathematics Courses Taken by Students**

|  |
| --- |
| Do more students enroll in 3rd and 4th year mathematics courses once the high school data science course is offered? Are course-taking patterns in 3rd and 4th year mathematics more diverse in student composition (comparing data science to other advanced math courses)? |

Course enrollment, achievement, demographic, and socioeconomic data for over 25,000 students dating back to 2018-19 from three school districts offering the data science course are used to investigate the relationship between the presence of the data science course and math class enrollment. First, a subset of 3,163 12th grade students who were enrolled 2021-22 and who had complete enrollment, achievement, demographic and socioeconomic data were used to investigate if taking data science was associated with the number of math classes students completed by the end of 2022.

A total of 73 data science and 3,090 non-data science students had complete data to conduct an Analysis of Covariance. The dependent variable is the cumulative number of math classes completed over the past 3 years including the Algebra I & II course series, Geometry, Pre-Calculus, AP Calculus, AP Stats, Other math classes, and Data Science. Covariates used to control for pre-existing differences enrolled and not enrolled groups include sex, age, indicators for race and ethnicity, years enrolled in the school, most recent algebra course grade, Free and Reduced Lunch (FRL) status, English Language Learner (ELL) status, Special Education (SPED) status, number of suspensions, and attendance rates.

Table 3 shows the demographic breakdown of the data science and non-data science student groups used for this analysis.

Table 3. *Demographic Characteristics of 12th Grade Students Enrolled and Not Enrolled in Data Science in 2021-22*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Statistic | Enrolled in Data Science | | Total |
| No | Yes |
| Sex (1 = female) | Mean (proportion) | .51 | .44 | .50 |
| Std. Dev. | .50 | .50 | .50 |
| Age in years  (as of May 31st, 2022) | Mean | 18.15 | 18.27 | 18.15 |
| Std. Dev. | .38 | .37 | .38 |
| White | Mean (proportion) | .49 | .66 | .49 |
| Std. Dev. | .50 | .48 | .50 |
| Black | Mean (proportion) | .01 | .00 | .01 |
| Std. Dev. | .12 | .00 | .11 |
| American Indian | Mean (proportion) | .07 | .22 | .07 |
| Std. Dev. | .25 | .42 | .25 |
| Asian | Mean (proportion) | .34 | .08 | .34 |
| Std. Dev. | .48 | .28 | .47 |
| Native Hawaiian | Mean (proportion) | .01 | .00 | .01 |
| Std. Dev. | .08 | .00 | .08 |
| Other Race | Mean (proportion) | .08 | .04 | .08 |
| Std. Dev. | .27 | .20 | .27 |
| Hispanic | Mean (proportion) | .30 | .51 | .31 |
| Std. Dev. | .46 | .50 | .46 |
| Free and Reduced Lunch | Mean (proportion) | .49 | .58 | .49 |
| Std. Dev. | .50 | .50 | .50 |
| English Language Learner | Mean (proportion) | .07 | .03 | .06 |
| Std. Dev. | .25 | .16 | .25 |
| Special Education Status | Mean (proportion) | .06 | .01 | .06 |
| Std. Dev. | .24 | .12 | .24 |
| Years enrolled | Mean (proportion) | 3.91 | 3.95 | 3.91 |
| Std. Dev. | .38 | .33 | .38 |
| Most Recent Algebra GPA | Mean | 2.81 | 2.52 | 2.80 |
| Std. Dev. | 1.20 | 1.06 | 1.20 |
| Number of Suspension | Mean | .00 | .00 | .00 |
| Std. Dev. | .10 | .00 | .10 |
| Attendance rate | Mean (proportion) | .91 | .93 | .91 |
| Std. Dev. | .17 | .05 | .17 |
| Semesters of math classes | Mean | 5.73 | 6.01 | 5.74 |
| Std. Dev. | 1.28 | 1.31 | 1.28 |

A series of tests of association reveal that the Data Science courses included students who were slightly older (*p* < .01), made up of more white students (*p* < .01), more American Indian students (*p* < .001), fewer Asian students (*p* < .001), more Hispanic students (*p* < .001), and students with slightly lower prior Algebra achievement (*p* < .05).

The ANCOVA reveals that, on average, the students enrolled in data science took 0.4 more semesters of mathematics over the past 3 years than students who did not take data science, after controlling for the differences in the set of covariates [*F*(1, 3146) = 10.73, *p* < .01][[2]](#footnote-2). In other words, almost every second student took an extra semester of mathematics over and above what students who did not enroll in data science ended up taking by the end of 2022. Figure 3 shows the results.

Figure 3. *Average number of math class semesters taken by 12th grade students after 3 years who were enrolled and not enrolled in data science*

*A picture containing text, screenshot, diagram, plot

Description automatically generated*

A similar analysis with a subset of 3,702 11th grade students of which 36 were enrolled in data science in 2021-22 shows a similar pattern of results but lacks statistical power to reveal significance.

Taken together the results show that a more diverse group of students take data science courses and they had slightly lower prior achievement than those not enrolled.

1. **Students Interest in STEM and views on Data Science course experience**

|  |
| --- |
| Do students who take the high school data science course see a future for themselves in STEM? Do students who take the high school data science course report that the learning is more meaningful and relevant than in previous mathematics courses? |

Students in two school districts were administered a survey measuring attitudes and beliefs about STEM and STEM related variables in the Fall of 2021 and Spring of 2022. A total of 841 students completed the survey in the Fall and 500 completed it in the Spring. Data collected in the Spring are used to answer the question of whether students who took the high school data science course see a future for themselves in STEM relative to their peers who had not taken the course. Three main variables are evaluated to answer the question including: student beliefs about having a future in STEM, interest in pursuing STEM, and confidence in data science skills relevant to STEM.

The belief about having a future in STEM is measured with a single self-report rating item: “*I know I have a future in STEM if I want one*”, where 1 = disagree and 10 = Agree.

Interest in pursuing STEM is measured using three 5-pt Likert items: “*How interested are you in pursuing a future career in STEM?*” (0 = not at all interested to 4 = very interested), “*How motivated are you to take STEM classes in college*”, and “*How motivated are you to major in a STEM field in college?*” (0 = ‘not at all’ to 4 = ‘a great deal’). The ratings for the three items are summed together into a single STEM interest score with a range from 0-12 and that has a Cronbach’s alpha reliability of 0.93.

Finally, confidence in data science skills relevant to STEM is measure with 4-items asking students to rate how confident they are in their skills to “*Use data to make a decision*”, “*Create a convincing argument using data as evidence*”, “*Assess the reliability of claims made in online sources (eg. social media, news sites, etc.)*”, and “*Interpret data presented in graphs, charts, and tables*”. Each item is rated with a scale from 1 to 10 and the sum of the four ratings is used as a single indicator of confidence. Scores range from 4 to 40 and the measure has a Cronbach’s alpha reliability of 0.90.

As a part of the analysis, three variables are employed to control for some of the potential pre-existing differences in the data science and non-data science groups of students. The variables are gender, support to pursue STEM, and grade. Gender is measured categorically with students indicating “male”, “female”, or “non-binary”. This variable is operationalized as two variables with the male category serving as the reference category in both instances. Support to pursue STEM is measured with the question “*Do you have people in your life that encourage you to learn STEM subjects?*”. Finally, grade is measured as their current self-reported grade (9, 10, 11, or 12).

Multivariate Analysis of Covariance (MANCOVA) is used to determine whether there are significant differences in the means of the data science and non-data science students on the three dependent variables while controlling for group differences in gender, having support to pursue STEM, and grade. This analysis is analogous to simultaneously conducting three independent sample t-tests. The MANCOVA reveals that data science students had significantly stronger beliefs they had a future in STEM if they wanted (*DiffDS – nonDS* = 1.23, *F*(1, 346) = 5.34, *p <* .05), more interest in pursuing STEM (*DiffDS – nonDS* = 1.59, *F*(1, 346) = 6.62, *p <* .05), and higher confidence in their data science skills relevant to STEM (*DiffDS – nonDS* = 5.51, *F*(1, 346) = 6.62, *p <* .01)[[3]](#footnote-3) Table 4 provides the estimated means and Figures 4 - 6 shows bar graphs that illustrate the results visually.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 4. *Mean Estimates for non-Data Science and Data Science Students* | | | | |
|  | Non-Data Science (n = 327) | | Data Science (n = 25) | |
| Mean | Std. Error | Mean | Std. Error |
| Future in STEM | 4.79a | 0.14 | 6.02a | 0.51 |
| Interest in STEM | 3.11a | 0.16 | 4.70a | 0.60 |
| Confidence in STEM Skills | 24.71a | 0.45 | 30.23a | 1.64 |
| a. Covariates appearing in the model are evaluated at the following values: female = .53, non-binary = .02, Do you have people in your life that encourage you to learn STEM subjects? = .30, What grade are you in? = 10.60.  Figure 4. *Students’ Expressing their belief that they can have a future in STEM, by course enrolment* | | | | |

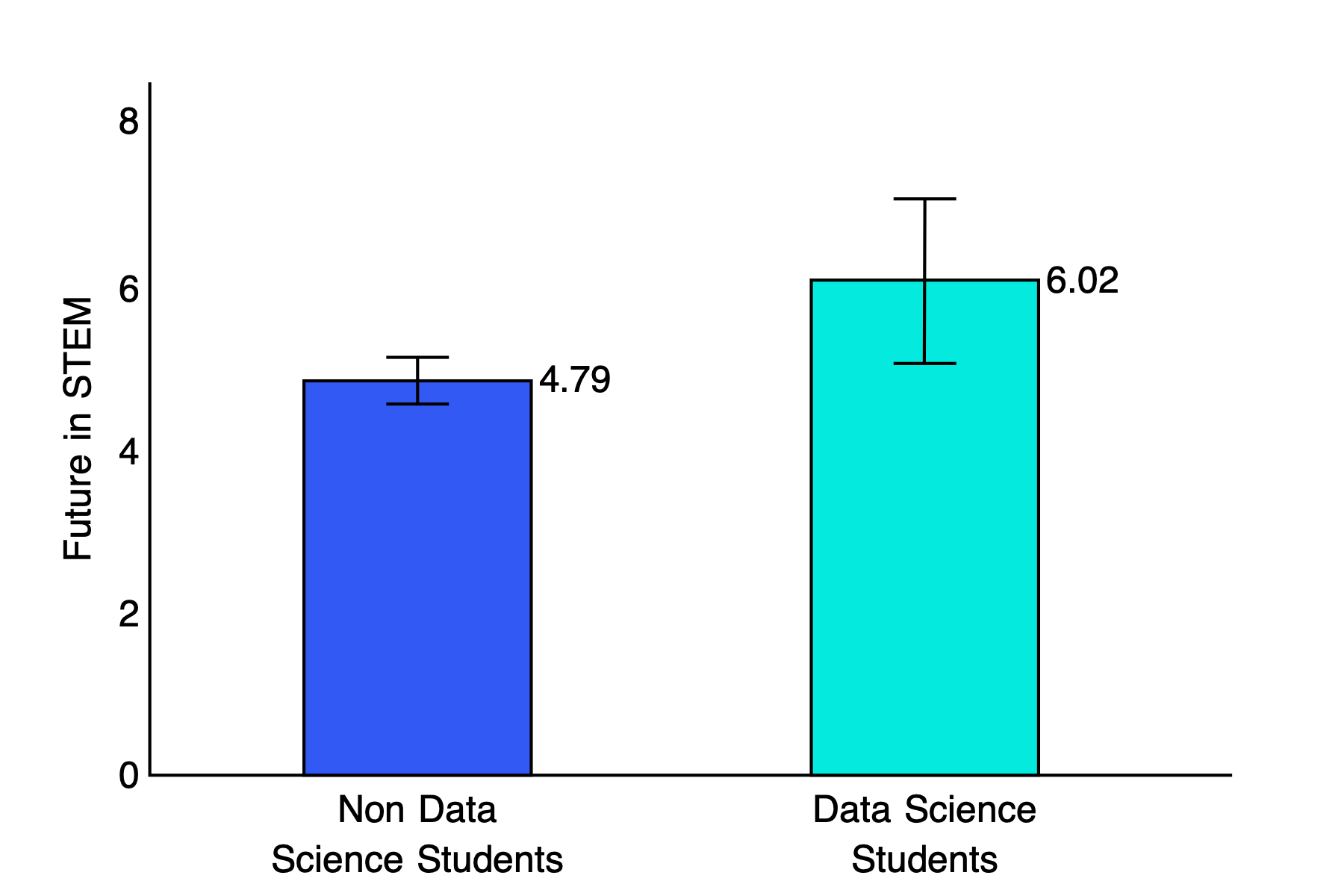


Figure 5. *Students’ Expressing their interest in STEM, by course enrolment*

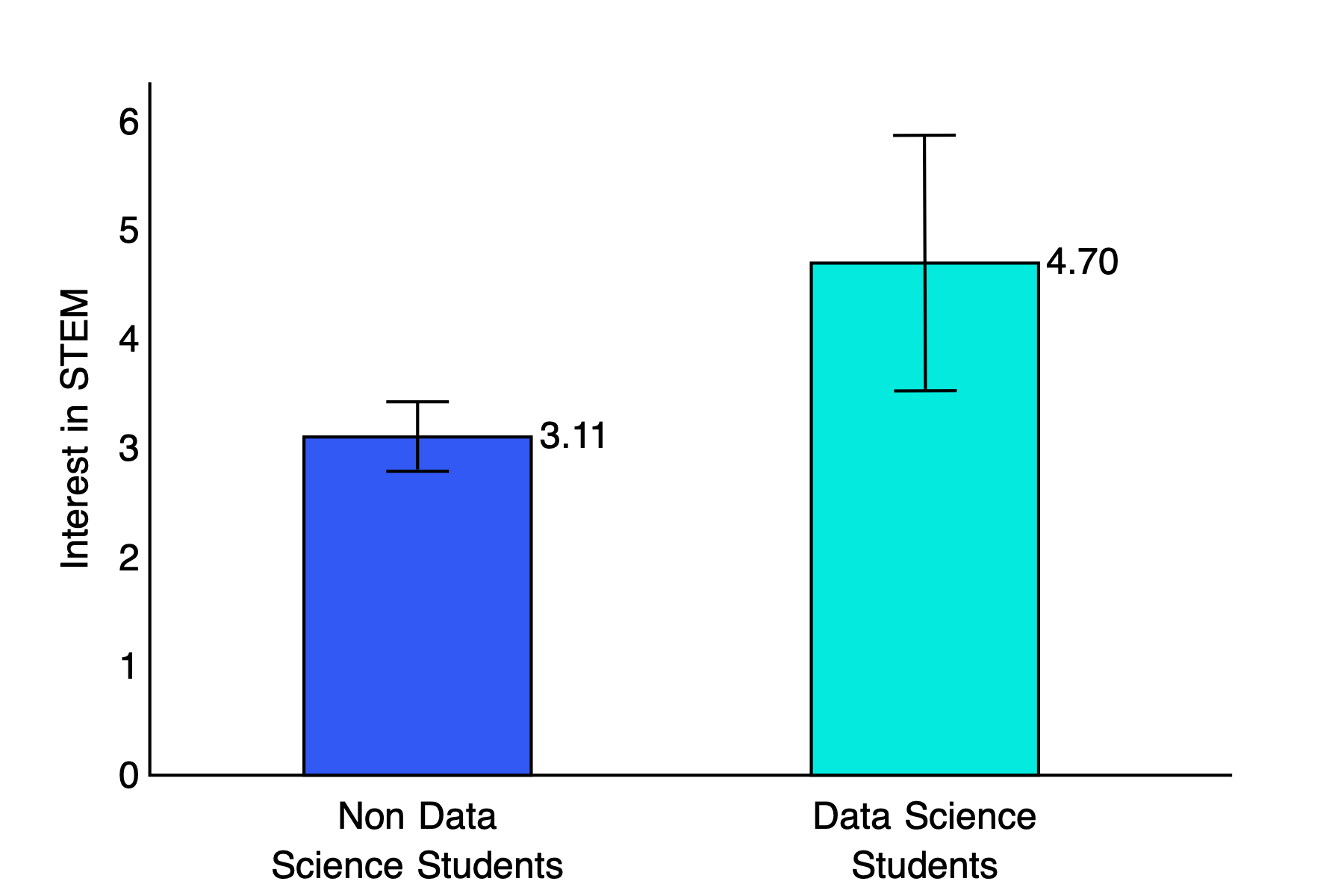
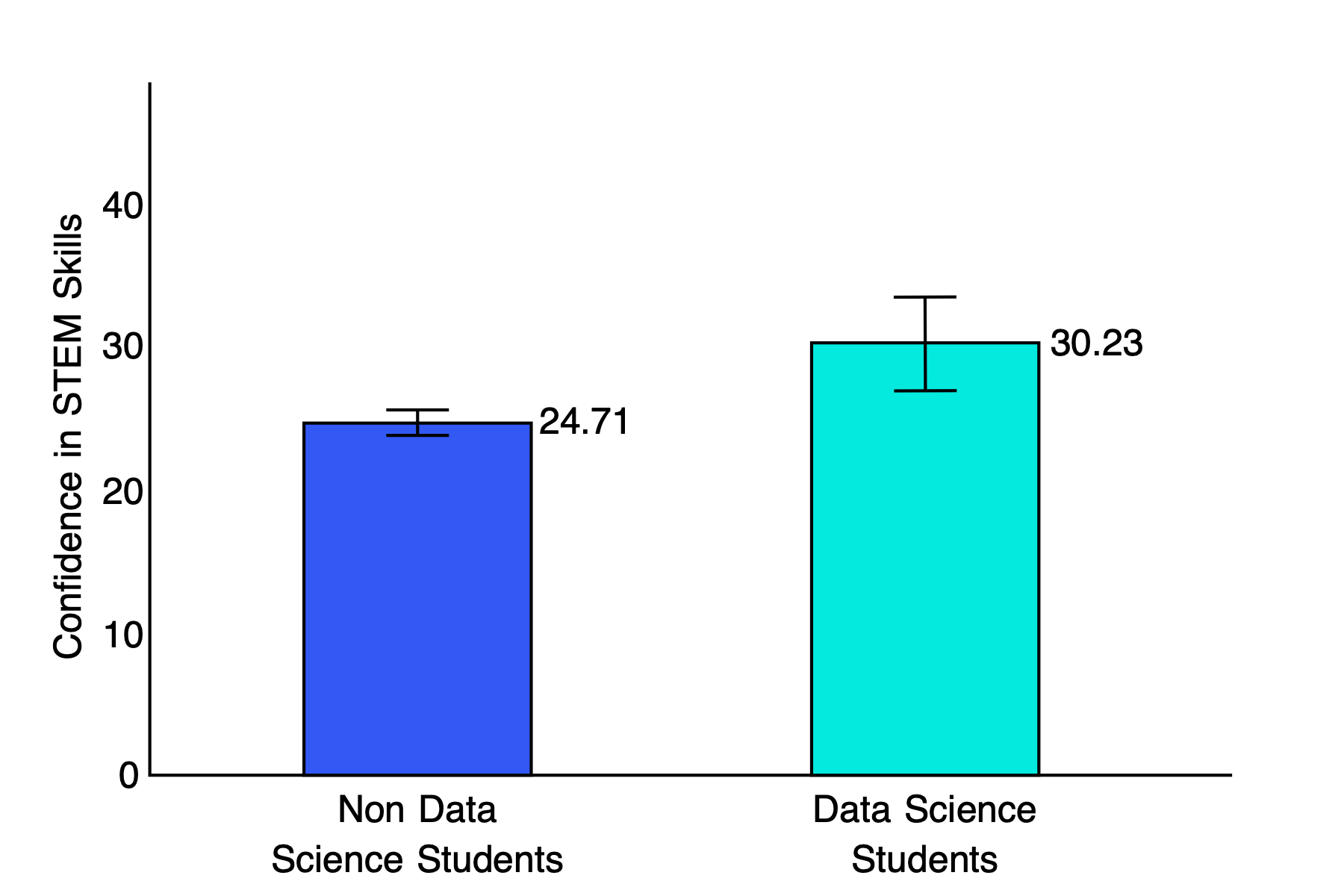


Figure 6. *Student confidence in their skills in data science and in judging the credibility of online information.*



Additional survey data collected in the Spring of 2023 from 11th and 12th grade students who had taken the Data Science course in eight high schools from the same two districts provides consistent evidence of a development of confidence in data science skills and interest in STEM. The mean total confidence score based on the same instrument described above is 27.42, (n = 323, 95% CI [26.68, 28.13]). In addition, 50% (150/301) and 36% (108/301) answered ‘yes’ to the questions “*Has this class increased your interest in pursuing a STEM Career?*” and “*Has this class increased your interest in pursuing higher education?*”

**Feedback from Students**

Quickwrites were administered to students via an anonymous survey given to students by their teachers. Students were given two prompts:

1. What are your thoughts about this data science course? What do you like, or not like, about it?
2. How would you describe this class to another student who is thinking about taking it next year?

The quickwrites were distributed roughly 2-4 months into the school year when students had completed the second unit of the course. Data has been analyzed for the responses from 138 students in 5 different classes. The responses were coded by a team of researchers to look for trends across students and classes.

Consistent themes emerged from the quickwrites. Students appreciated the connections in class to real-world data and ideas. They appreciated the opportunity to engage in discussions, and the opportunity to learn how to interpret data and understand the stories it tells. Students were positive about the interdisciplinary nature of the course and the fact it challenged them to integrate creative thinking, writing and mathematical reasoning. The students also reported their appreciation for learning practical skills such as fluency with computer programs (for example Python, Codap, Google Colab, Google sheets, Tableau) and tools that could be relevant for their lives and for future careers. Students described the connections they could see between their learning of data science and their future career opportunities beyond school.

Students also reported enjoying the hands-on activities and the fact that the course was project-based without tests, which reduced stress and improved their learning experience. They appreciated the diversity in the homework that was assigned such as completing brief reflections and collecting data. Students conveyed gratitude for the collaborative and interactive nature of the course with its emphasis on group work and discourse. The students found the course to be different from their previous experiences of mathematics – not in the level of rigor or the difficulty of the concepts – but in the ways students could access the concepts and in the diversity of the ideas they encountered. The students reported that their previous experience of mathematics was one of right and wrong answers, which contrasted with their learning of data science, which was more typically about the use, application, and interpretation of a broad set of ideas.

Overall, students expressed very positive views of the course. Some students did express that they disliked or struggled with the coding aspects of the course or that they felt certain assignments could be clearer - suggestions for improvements were passed onto the writing team. Specific student critiques were noted, and where relevant, will be incorporated into continuing refinement of course materials.

Feedback from Teachers

Interviews were conducted with multiple teachers who taught the data science course, across the research sites. One of the questions asked teachers how they would respond to critiques that data science is a lower level course, not appropriate for students going into STEM. Two of the responses are below:

*1.If you want to look at calculations, there's not a lot of calculating itself [in the Data Science Course]. So if that's all they're looking at, then I can see that perspective. But if you want to look at the level of the thought process that needs to go into it, I think it's a really high thought process class about how much are you thinking about it, how diverse are you thinking about it, what are all the parameters that could be pulled for it? It can get overwhelming because, to me, it's such a high thought level and it's really, I think for my group, I'm just starting to see them.*

*The Data Science students thought process just kind of emerged on their own without a whole heck of a lot of prompting from me. And I think they're really growing in that and that's going to benefit them later. And I don't think there's a lot of classes at the high school level that lets them have that open, analytical, push that higher, thought process into a lot of different areas and avenues. And for that, I don't think people realize that's what this class is allowing to have happen. And I think it's really needed for the kids.*

2.*I think it [Data Science course] prepares students in a different way. So I think it actually teaches them much more of the scientific thinking process. I think it teaches them how to acquire new skills that are necessary to become someone that does STEM. So I think what it's teaching is actually more the thinking skills and the practical day-to-day operations of someone who works in the STEM field. So I think that's what's good.*

*3.There are different sets of knowledge…. It's not a low level math class. It's a little bit like comparing apples and orange*s…*.And prior to being a student teacher, I worked as a software engineer in San Francisco, and there was a data science team, and that data science team used Tableau [technology tool] for what they were doing. So that was just like, oh, that's awesome. This is something that is actually being used in the real world that students are learning about. So, I just wanted to point out that there's real world application to the stuff going on that I can clearly see.*

**CONCLUSION**

Our results, and [other studies](https://edpolicyinca.org/publications/innovating-high-school-math-through-k-12-and-higher-education-partnerships), show that data science provides an important high school course option for students. It teaches students important mathematics, statistics, and data, in ways that they can use the ideas, achieving at higher levels than students learning algebra 2. The course also means students take more mathematics and they are encouraged to go further in STEM and in higher education.

A great challenge to the course being taken up in larger numbers is the growing movement of opposition, with traditionalists working to dial back UC decisions to encourage data science in admissions. This is a growing movement, making research results on courses such as this one, all the more important.

An important part of the strategy in preparing other districts to adopt and implement data science is clear communication to districts, countering misinformation. Information on data science in the education system and its importance for students’ futures is critical for teachers, parents and leaders in districts. All of the districts in this study were given parent information evenings.

Parents, teachers and students are excited about the content of data science, and students leave the course empowered to read and understand the word, and go further in a STEM journey. It will be the responsibility of researchers to continue to conduct research on data science courses, and to communicate as widely as possible about the role data science plays in countering inequities in STEM.

1. The analysis also included school district as a factor that might interact with whether students were enrolled in Data Science to predict performance. Results revealed students in one of the districts performed significantly better than in the other but that this did not interact with whether students were enrolled in Data Science. Hence, only the main effect of the Data Science course is reported. [↑](#footnote-ref-1)
2. The assumption of equality of error variance for the dependent variable is met. The normality assumptions for the two cells of the design were not but results from ANOVA tend to be robust to violations of non-normality. [↑](#footnote-ref-2)
3. Multivariate outlier analysis resulted in the removal of two non-Data Science student cases. Assumptions of equality of covariance across the dependent variables for the two groups and equality of error variance for each dependent variable are met. While the normality assumptions of all cells of the design for each dependent variable were not met, however results from MANOVA are robust to violations of normality in the data [↑](#footnote-ref-3)