ImpactEd Evaluation

Science Stars

IMPACT EVALUATION REPORT

January 2026

ImpactEd Evaluation

Contents

Executive Summary

Science Stars is a widening participation tutoring intervention delivered by City St. George's, University of London. It aims to improve the science GCSE attainment of Year 11 pupils. The academic year 24/25 was the sixth year of the Science Stars tutoring programme and this evaluation has found potential benefits on pupils' non-cognitive and social, emotional and mental health (SEMH) outcomes including pupils' metacognition, self-efficacy and test anxiety.

Following the success of last year's delivery, the programme was delivered in-person at the same two schools: Burntwood and Ernest Bevin. These are both single- sex schools; Burntwood is all-girls whilst Ernest Bevin is all-boys.

This evaluation used a mixed-methods approach, combining quantitative attainment and non-cognitive surveys with qualitative interviews and focus groups. The survey and attainment data was collected for a participating group and a comparison group for each school. The comparison group was selected by finding a cluster of students from the same school, same year group and similar target grades (where possible) as the Science Stars participants. The qualitative work was carried out with those delivering the Science Stars programme (students from the university known as student ambassadors) and 1-1 interviews with the teachers co-ordinating the programme in each of the schools.

Analysis of findings from the first two years of the evaluation of the Science Stars programme (2019/20 and 2020/21) showed positive impacts of the programme; the next year (2021/22) showed the beginning of some negative trends. Yet progressively, the evaluation of the last two years (2022/23) and (2023/24) showed positive results. This year (2024/25) explores some mixed findings.

Overall, analysis of outcomes showed that participating pupils improved their non-cognitive and SEMH skills compared to a comparison group. Some of the positive trends from last year have continued, such as an improvement in pupils' self-efficacy. Yet, for some outcomes, including change in GCSE grades, the impact is more modest when compared to the comparison group.

Key findings

Overall, participants demonstrated greater improvements in non-cognitive skills than their comparator peers, though comparator peers showed better attainment outcomes. These findings are summarised below.

- On average, participants' metacognition increased (+0.21 percentage points) while their comparator peers' metacognition decreased (-6.63 percentage points), suggesting the Science Stars programme positively impacted participants' revisions skills and understanding of their own learning. This finding was statistically significant (p=0.02).
- On average, participants improved their self-efficacy (+5.49 percentage points) while their comparator peers' self-efficacy decreased (-19.79 percentage points), suggesting Science Stars had a positive impact on participants' confidence in science. This finding was statistically significant (p=0.01).
- On average, participants' levels of test anxiety decreased (-1.64 percentage points) while their comparator peers' test anxiety rose (+5.03 percentage points), suggesting that Science Stars may have had a positive impact on participants' anxiety around tests and exams. This finding was not statistically significant (p=0.24).
- The difference between participants and comparator peers across all three measures at Ernest Bevin is more positively pronounced than the overall trend, and the differences are statistically significant for metacognition (p=0.01) and self-efficacy (p=0.01).
- Burntwood's participants fared worse than their comparator peers in their metacognition and test anxiety. These differences, however, were not statistically significant (p=0.07 and p=0.70 respectively), meaning this could be due to random chance rather than a true effect of the programme.
- On average, participants increased their science grade from 4.30 to 4.44 from their Spring mocks to their final GCSE grade (+1.79 percentage points) which was less than the progress made by comparators (3.19 to 3.83, +7.93 percentage points). This difference was not statistically significant (p= 0.053).
- 74% of participants achieved their target grade, which was slightly less than the 77% of comparator peers who reached their target. This difference between participants and comparators was not statistically significant (p=0.82).
- Pupils from Ernest Bevin showed a larger increase (13.6 percentage points) than their comparator peers (9.38 percentage points) from their Mock grade in the Autumn term to

their final GCSE grade in September 2025. This difference was not statistically significant (p=0.055).

- 94% of participants at Ernest Bevin achieved their target grade compared to 100% of their comparator peers. This was sixteen out of the seventeen participants that achieved their target grade. This difference between participants and comparators was not statistically significant (p=0.36).
- A greater percentage of Burntwood participants (56%) achieved their target grade in comparison to their comparator peers (45%). This was not statistically significant (p=0.61).

Longitudinal Findings

- This year saw a continuation of last year's trend; participants saw positive progress in metacognition, test anxiety and self-efficacy. The positive progress in self-efficacy was larger than the previous year, while the progress in metacognition and test-anxiety was smaller.
- The percentage of participants achieving their target grade decreased by 8.57% from last year to this year.
- A smaller proportion of participants (74%) achieved their target grade compared to comparators (77%). This difference between groups did not reflect the positive difference from previous years.

Recommendations

Evaluation Recommendations

- Investigate external factors that may have affected pupils' lower gains in metacognition and test-anxiety compared to the previous year. This could involve further exploration of tutoring strategies and ability to tailor learning to the pupils' needs.
- Consider a more rigorous matching procedure such as propensity score matching so that more robust conclusions can be drawn between participating and comparator pupils.
- Consider exploring pupil voice where possible to triangulate findings.
- Continue to increase sample size where possible. Increased sample size will improve the robustness of the findings. It could be worth exploring whether the programme could be scaled and retain the same benefits with a larger group.

Programme and Delivery Recommendations

- Consider sharing further details on pupils' academic abilities with the student ambassadors prior to the tutoring sessions starting.
- Consider strategies to monitor the attendance of the pupils.
- Consider an exploration of teaching and learning styles prior to the programme starting so that sessions can be tailored to meet individual needs.
- Ensure feedback to schools is always timely so that those involved can be addressed prior to the next session.
- Make sure all data collection is consistent across schools so that comprehensive data is shared to strengthen findings.
- We recommend investigating what Ernest Bevin was doing particularly well to lead to such positive results, and exploring how this could be emulated across other schools

Introduction

In 2024 St George's, University of London merged with City, University of London to form City St Georges, University of London. City St George's is now a large multi-faculty institution with a focus on professional education and research. The combined university is now one of the largest suppliers of the health workforce in the capital and one of the largest higher education destinations for London students. With a strong historical commitment to widening participation activities, City St George's is increasingly working across the whole student lifecycle to support students from under-represented backgrounds. This year, City St George's has run the Science Stars programme for a sixth year, focusing specifically on school-based activities to raise attainment.

ImpactEd Evaluation is a not-for-profit organisation that exists to improve pupil outcomes by addressing the evaluation deficit in education. ImpactEd Evaluation works in partnership across the education sector to support high-quality monitoring and evaluation that informs decisions about what will work most effectively to support students. Their work in access and widening participation has included evaluation projects with University College London, Goldsmiths University and London South Bank University, among others.

Programme Overview

Science Stars is a widening participation programme aimed at improving educational outcomes specifically for students who already face additional barriers in their education. While participants are not required to match a widening participation criterion, schools were asked to prioritise pupils who are in receipt of pupil premium. Science Stars is a sustained tutoring intervention designed to support Year 11 pupils to prepare for GCSEs and ultimately increase their attainment in science. The programme is delivered in-person by current students at City St George's, University of London, referred to in this report as student ambassadors, following a pre-designed curriculum developed by a former science teacher.

The programme was implemented at two single-sex schools: Ernest Bevin (all-boys) and Burntwood (all-girls). This gender composition is an important contextual factor when analysing differences in both non-cognitive outcomes and academic attainment between the schools.

The programme aims to improve educational outcomes in GCSE Science for target pupils in Year 11. The key aims and objectives of the programme for participating pupils are as follows:

- More able to answer exam questions.
- Better understanding of science GCSE content.
- Increased academic attainment.
- Improved revision skills.
- Better understanding of their own learning, strengths, and weaknesses.
- Increased confidence in science.
- Less anxious about tests and exams

Methodology

Research Questions

The evaluation aims to answer the following three research questions:

- What is the impact of the Science Stars programme on pupils' confidence in science?
- What is the impact of the Science Stars programme on pupils' revision skills, test anxiety and understanding of their own learning, strengths and weaknesses?
- What is the impact of the Science Stars programme on pupils' attainment in Science GCSE?

Outcome Measures

The table below shows the key outcomes in this evaluation and how they will be measured using both quantitative and qualitative measures.

Table 1 shows the key outcomes and measures

Outcome	Quantitative Measure	Qualitative Measure
Improved revision skills	MSLQ Metacognition	Focus groups with student ambassadors (those at the university that run the sessions with the students) and interviews with teachers (those that co-ordinate the sessions).
Increased confidence in science	MSLQ Self-efficacy	
Less anxious about tests and exams	MSLQ Test Anxiety	
More able to answer exam questions	School attainment data	
Better understanding of their own learning, strengths, and weaknesses	MSLQ Metacognition	
Better understanding of science GCSE content	GCSE grades and school attainment data	
Increased academic attainment	GCSE grades and school attainment data	

Evaluation Design

This evaluation is the sixth annual evaluation of this programme and it was conducted over the academic year 2024/25. All of the data was collected between Autumn Term 2024 and Autumn Term 2025.

As pupil selection was conducted by the school and through a voluntary sign-up process, a randomised control group design was not possible. As such, a comparison group was formed by finding a cluster of students from the same school, same year group and similar target grades (where possible) as the Science Stars participants. Including a comparison group allows us to assess how participating pupils performed on relevant outcomes relative to a similar group of non-participating peers, helping us to isolate the effect of the programme, understand its impact on their non-cognitive outcomes and attainment, and identify areas for improvement. This group will be referred to as the comparator group throughout the report. It should be noted that due to the simplified matching approach, comparisons drawn between the two groups should be interpreted with appropriate caution, as it may not account for all relevant underlying factors. For any future evaluations, City St George's, University of London should consider using propensity score matching for a more robust approach for selecting a control group. In addition, due to this selection, it is likely that participants in the intervention group are more likely to be facing additional barriers to learning than the comparator group so this should be considered while exploring findings in this report.

Although there are some limitations to this design approach (referenced in the 'Limitations' section of the methodology), it allows us to make relatively robust inferences within these constraints by collecting a range of datapoints to triangulate findings and assess if there was a common pattern across indicators.

In this evaluation, we analysed three different types of data:

- **Pupil survey data** was used to evaluate the impact of the programme on pupils' non-cognitive outcomes, such as metacognition, self-efficacy and pupils' test anxiety.
- **Attainment data** was used to evaluate the impact of the programme on pupil's academic progress in science.
- **Qualitative research** was used to evaluate the success of the implementation of the programme as well as its impact on pupils.

Survey: Design and Sample

The non-cognitive outcomes (self-efficacy, test anxiety and metacognition) were measured because they have predictive validity, i.e., they have been shown to be associated with improvements in long-term outcomes such as well-being, academic achievement, and employment destinations. Alongside academic achievement, there is evidence that these skills can be particularly important in closing disadvantage gaps.

These non-cognitive outcomes were measured using psychometrically validated questionnaires, administered to pupils pre and post Science Stars. Pupils were assessed at the beginning (baseline collection) and end (final collection) of the programme. Collecting data at these two time points allows us to analyse the level of change over the course of the programme for each specific outcome.

Our core outcome measures for this evaluation were:

Table 2 provides details on the outcome measures

Outcome	Measurement Details
Metacognition	Metacognition means 'thinking about thinking': pupils' ability to think explicitly about their own learning. It is strongly associated with academic progress and improves other skills required for learning, such as critical thinking (Flavell, 1979; Higgins et al., 2016). We measured metacognition using the Cognitive Strategies Use and Self-Regulation subscales of the Motivated Strategies for Learning Questionnaire.
Self-efficacy	Self-efficacy is a measure of pupils' belief in their ability to achieve a specific task in the future. Self-efficacy is correlated with higher academic achievement and persistence, and also contributes to pupil wellbeing (Gutman & Schoon 2013, DeWitz et. al. 2009). We measured self-efficacy using the Self-efficacy subscale of the Motivated Strategies for Learning Questionnaire.
Test anxiety	Test anxiety is concerned with pupils' emotional responses to tests (Pintrich and De Groot, 1990). Greater levels of test anxiety can result in worse performance in exams but in some situations may be linked to increased motivation.

The results of the psychometrically validated surveys will be supplemented by qualitative data from two focus groups with five Science Stars student ambassadors and two one-to-one interviews with the teachers, one from Burntwood and the other from Ernest Bevin.

The table below summarises what surveys were completed, who responded, and the sample size of respondents.

Table 3 shows the sample size for surveys

Data	Which pupils?	Matched Sample Size		
		Ernest Bevin	**Burntwood**	**All pupils**
Meta-cognition	Participating	12	13	25
	Comparator	15	8	23
Self-efficacy	Participating	14	13	27
	Comparator	15	9	24
Test anxiety	Participating	14	14	28
	Comparator	15	9	24

Attainment data: Design and Sample

The table below shows what attainment data was collected, when it was collected, whose attainment data was collected, as well as the sample size. Burntwood did not share Autumn Mock data.

Table 4 shows the sample size for attainment

Data	When?	Which pupils?	Matched Sample Size		
			Ernest Bevin	**Burntwood**	**All pupils**
Target grades	Autumn Term 2024	Participating	17	18	35
		Comparator	15	11	26

Autumn Mock exam	Autumn Term 2024	Participating	17	0	17
		Comparator	15	0	15
Spring Mock exam	Spring Term 2025	Participating	17	18	35
		Comparator	15	11	26
Final GCSE results	September 2025	Participating	17	18	35
		Comparator	15	11	26

Qualitative Research: Design, Sample and Analysis

Focus groups were conducted with the student ambassadors leading the tutoring sessions at both schools. Three student ambassadors participated in a focus group for Ernest Bevin, and two student ambassadors participated in a focus group for Burntwood. 1:1 interviews were conducted with the relevant teacher co-ordinating the sessions in each of the schools.

The qualitative data was analysed using a deductive thematic approach, meaning that we systematically 'code' the data to find common themes and present these, drawing on examples where appropriate.

Limitations

There are some limitations of this evaluation design worth noting:

- The comparison group was not created through random allocation. This means there may be unmeasured differences between the intervention and comparison groups (beyond prior attainment such as student motivation or parental support) that could have affected pupil outcomes.
- Particularly when analysing the schools separately, the overall sample size for both participants and the comparator group is small. As such, results may not be immediately generalisable to other school contexts.
- It was noted by City St George's that some participants at Burntwood started the programme late and only received two to five sessions. Therefore, it is important to consider this context when analysing the results.
- This report examines performance differences between two schools: one with seven years of programme experience (Ernest Bevin) and another in its second year of implementation (Burntwood). Therefore, implementation and experience should be considered when studying the outcomes.

- Most of the time, qualitative analysis is used to qualify and to explain differential impact on the two schools participating in the programme. Feedback here was limited due to the low number of student ambassadors who participated in the focus groups.
- Autumn term attainment data was only collected for one school. Consistency between schools would encourage more robust findings.
- Although pupil outcomes are a key focus of outcomes, no qualitative research with pupils was conducted as part of this evaluation. City St George's, University of London, could consider conducting qualitative research with pupils to triangulate some more of the quantitative findings, such as the impact of the programme upon pupils' understanding of their own strengths and weaknesses, their test anxiety, confidence and science attainment.
- Statistical significance: When reporting on statistical significance, we use the standard social science convention of a 'significant' probability value (p-value) being less than 0.05. This means that the likelihood of observing changes at least as severe as those observed, if it were, in fact, the case that the intervention had no impact, is less than 0.05 (i.e. highly improbable). If a finding is not statistically significant, this does not rule out an effect, but it means that we cannot confidently say that the changes observed were not due to random chance.

Outcomes 24/25

Non-cognitive and SEMH skills- all pupils

Key finding: On average, participants' metacognition increased (+0.21 percentage points) while their comparator peers' metacognition decreased (-6.63 percentage points), suggesting the Science Stars programme positively impacted participants' revision skills and understanding of their own learning.

The difference between participants and comparator pupils was statistically significant (p=0.02), suggesting the change is less likely to be due to chance. Despite the positive change being relatively small it suggests that the Science Stars programme enabled pupils' to maintain their revision skills and understanding of their strengths and weaknesses in comparison to their peers, whose findings show a negative change in their metacognitive skills.

Table 5 shows participant and comparator findings for metacognition

	Type of Pupils	Sample size	Baseline	Endline	Percentage point difference	Statistical significance
Metacognition	Comparators	23	4.16	3.77	-6.63	*p* = 0.02
	Participants	25	4.36	4.37	0.21	

An improvement in revision skills and pupils' understanding of their own learning was also recognised by the student ambassadors. They noticed that the practical questions in the sessions helped pupils to develop their understanding while seeing the key concepts put into practice. One student ambassador said drawing out the concepts helped the pupils to visualise them in their heads and asking them to explain the concepts in different ways helped the pupils to recognise their own strengths and weaknesses:

"I think that was a good exercise using the exam questions, but also just verbally saying right, if I came down from Mars, how would you explain this concept to me and things like that. I think they suddenly would sometimes realise, oh, I don't know it or I do."

- *Student Ambassador*

Key finding: On average, participants improved their self-efficacy (+5.49 percentage points) while their comparator peers' self-efficacy decreased (-19.79 percentage points), suggesting Science Stars had a positive impact on participants' confidence in science.

The difference between participants and comparator pupils' self-efficacy scores was statistically significant (p=0.01). This suggests that this finding is less likely to be due to chance. The large difference between participants and comparators is certainly compelling because the comparator groups confidence declined over the same period suggesting the Science Stars programme protected participants from this decline and boosted their confidence beyond the starting point.

Table 6 shows participant and comparator findings for self-efficacy

	Type of Pupils	Sample size	Baseline	Endline	Percentage point difference	Statistical significance
Self-efficacy	Comparators	24	4.70	3.47	-19.79%	*p* = 0.01
	Participants	27	4.65	5.02	5.49%	

This increase in confidence was supported by student ambassadors and teachers at both schools. A student ambassador at Burntwood stated that the pupils became ***"a lot more confident to participate and also just to say when they didn't understand something".*** This statement also emphasises the pupils' enhanced understanding of their strengths and weaknesses throughout the course of the programme. Similarly, the student ambassador at Ernest Bevin said that they noticed an increase in participation from quieter pupils over time suggesting that they had more confidence to approach questions even if they were unsure of the answer.

The teacher from Burntwood leant into the relationships with the tutors and how this may have supported pupils' confidence:

"I think some of them had a really good bond with their tutors, so they were open to saying, look, I think I am struggling. I don't think this is just the knowledge. I think I might need some extra time."

- *Teacher, Burntwood*

This teacher also noted that colleagues in the department had observed increased pupil confidence when approaching exam questions, particularly in their willingness to attempt questions and tackle challenging problems.

The teacher from Ernest Bevin said:

" It's amazing to see that these kids have such confidence in themselves.."

- *Teacher, Ernest Bevin*

This teacher stated that throughout the year there seems to be a transition point, where pupils go from thinking Science Stars will help them to just pass, to later thinking that they will get a good grade. The teacher mentioned that pupils had commented on finding foundation papers easy later on in the year and said that they would often ask what they needed to do to take the higher paper, suggesting a change in their self-efficacy over time.

Key finding: On average, participants levels of test anxiety decreased (-1.64 percentage points) while their comparator peers' test anxiety rose (+5.03 percentage points), suggesting that Science Stars may have had a positive impact on participants' anxiety around tests and exams. This was not statistically significant (p=0.24).

The difference between participants and comparator pupils was not statistically significant. This means we cannot be confident that Science Stars caused the differences observed—they may simply be due to chance.

Table 7 shows participant and comparator findings for test anxiety

	Type of Pupils	Sample size	Baseline	Endline	Percentage point difference	Statistical significance
Test anxiety	Comparators	24	4.8	5.1	5.03	P=0.24
	Participants	28	4.5	4.4	-1.64%	

Non-cognitive and SEMH skills- Ernest Bevin

Key finding: The difference between participants and comparator peers across all three measures at Ernest Bevin is more positively pronounced than the overall trend, and the differences are statistically significant for metacognition and self-efficacy.

The data from Ernest Bevin shows a similar picture to the overall positive trend described above; participants outperforming comparator pupils in both non-cognitive skills (metacognition and self-efficacy) and the one SEMH measure (test anxiety).

Table 8 shows participant and comparator findings for non-cognitive outcomes for Ernest Bevin

Non-cognitive skill	Type of Pupils	Sample size	Baseline	Endline	Percentage point difference	Statistical significance
Metacognition	Comparators	15	4.20	3.43	-12.84	P=0.01
	Participants	12	4.25	4.34	1.47	
Self-efficacy	Comparators	15	4.75	2.79	-32.59	P=0.01
	Participants	14	4.73	5.13	6.61	
Test anxiety	Comparators	15	4.73	5.33	10.00	P=0.07
	Participants	14	3.70	3.48	-3.57	

The change in self-efficacy was recognised by one student ambassador from Ernest Bevin who said:

"...walking through those scary looking questions, they slowly gain confidence in approaching them. So by halfway through the programme, they weren't as hesitant to approach it themselves."

- *Student Ambassador, Ernest Bevin*

In addition, the teacher at Ernest Bevin recognised a difference between Science Stars pupils and others within the school:

"They're more anxious at the very, very beginning. But then in the middle section where everyone else is running around like headless chickens, they were calm and relaxed."

- *Teacher, Ernest Bevin*

This teacher had also noticed a difference in Science Stars pupils' confidence in science lessons because they were able to access the learning more easily and they showed a better sense of their own understanding of science.

The student ambassadors also noticed a heightened willingness to complete questions in test conditions towards the end of the programme, suggesting less anxiety around tests and exams. In addition, the teacher explained that the pupils began to take their learning more seriously which was shown through asking for past papers, video links and resources to support their learning further.

Non-cognitive and SEMH skills- Burntwood

Key finding: Burntwood's participants fared worse than their comparator peers in their metacognition and test anxiety. These differences, however, were not statistically

significant, meaning this could be due to random chance rather than a true effect of the programme.

The broad picture for Burntwood pupils' non-cognitive and SEMH skills deviates from the overall positive trend initially described. Participants metacognition slightly decreased (-0.95 percentage points), while the comparator group showed an improvement (+5.02 percentage points). This change was not statistically significant (p=0.07). Participants self-efficacy increased (+4.27 percentage points) which was a greater positive change than their comparator peers (+1.54 percentage points); this difference was not statistically significant (p=0.60). Participants' test anxiety went up slightly (+0.30 percentage points) whereas the comparator groups' test anxiety decreased (-3.24 percentage points). The difference between these two groups was not statistically significant (p=0.70).

Table 9 shows participant and comparator findings for non-cognitive outcomes for Burntwood

Non-cognitive skill	Type of Pupils	Sample size	Baseline	Endline	Percentage point difference	Statistical significance
Metacognition	Comparators	8	4.09	4.39	5.02	P=0.07
	Participants	13	4.46	4.40	-0.95	
Self-efficacy	Comparators	9	4.49	4.59	1.54	P=0.60
	Participants	13	4.66	4.91	4.27	
Test anxiety	Comparators	9	4.78	4.58	-3.24	P=0.70
	Participants	14	5.25	5.27	0.30	

The self-efficacy of the pupils certainly reflects the positive findings of the 24/25 cohort with a positive change (+4.27 percentage points) which was larger than the comparator peers (1.54 percentage points). Even though this was not statistically significant it reflects similar findings to other pupils who took part on the programme suggesting that the Science Stars programme could have supported this increase in confidence.

Findings from the qualitative work with Burntwood student ambassadors and the teacher also supported these findings. Student ambassadors at Burntwood reported an increase in pupils' self-efficacy, observing greater confidence in providing verbal responses and reduced concern about making errors as the sessions progressed. In addition, the ambassadors for Burntwood felt that they were able to help the pupils develop their exam techniques which supported the pupils' approach towards tests and exams:

“I think towards the end they knew how to say things as the examiner would want to hear, which was really good and they could use keywords they were unfamiliar with and integrate them into their answers."

- *Student Ambassador, Burntwood*

The teacher from Burntwood also commented upon the organisation and delivery of the sessions:

"We had a couple [of students] that also have English as an additional language, so I think having a smaller group encouraged them to verbalise, which was really beneficial for them…I think all of them struggle with using those key terms. But if they're hearing it [in these sessions], they ask questions."

- *Teacher, Burntwood*

This suggests that the difference in pupils' approaches to learning as well as the student ambassadors' approach to tutoring could potentially account for some of the differences in scores. This may be something that Science Stars would want to explore further, looking particularly at the implementation of the tutoring sessions but also the preferred learning techniques of the pupils. This exploration may offer some consideration to the approach of the tutoring for different types of learners.

Attainment- All pupils

Key finding: On average, participants increased their science grade from 4.30 to 4.44 from their Spring mocks to their final GCSE grade (+1.79 percentage points) which was less than the progress made by comparators (3.19 to 3.83, +7.93 percentage points). This difference was not statistically significant (p= 0.053).

While participants made progress from their mock grades in the Spring term, the comparison group demonstrated larger gains over the same period. This was not statistically significant which means we cannot be confident that Science Stars caused the differences observed- they may be simply due to chance. It is important to note here that the comparator group on average started with lower Spring Mock scores suggesting there was more room for improvement than those starting at higher levels in the participant group. Further analysis of programme implementation could help to unpick why participants made less progress than their comparator peers.

Table 50 shows attainment findings for participant and comparator pupils

Attainment	Type of Pupils	Sample size	Spring grade	Endline - GCSE	Percentage point difference	Statistical significance
Science	Comparators	26	3.19	3.83	7.93	$p = 0.053$
	Participants	35	4.30	4.33	1.79	

Key finding: 74% of participants achieved their target grade, which was slightly less than the 77% of comparator peers who reached their target. This difference between participants and comparators was not statistically significant ($p = 0.82$, n = 61).

The majority of both participants and their comparator peers achieved their target science grades. While the comparator group had a marginally higher success rate, this difference was not statistically significant. This indicates that the difference may be better explained by random chance rather than representing a genuine programme effect. In addition, in the interviews with teachers and through communications with City St George's, information was shared on how target grades were set across schools. In Ernest Bevin target grades were decided internally within the school. They look at their KS2 data and other grades achieved leading up to year 11 and use those as a predictor. They also look at other pupils achieving similar grades to support their predictor. In Burntwood, however, the target grades were generated from a baseline assessment that was completed in Year 8. This inconsistency between the schools could explain the some of the differences in outcomes.

The range of target grades for the 35 participants was 3.5, the mode was 4 and the median was 4 while the range for the comparators was slightly larger at 4, with the mode as 4 and the median as 3.5. This suggests that the two groups are broadly similar overall in terms of their target grades, with both groups most commonly being targeted at a grade 4. However, the participants appear to be very slightly higher attaining on average, with a median target grade of 4 compared to 3.5 for the comparators, and a slightly narrower range suggesting their grades were somewhat more clustered together. This is worth bearing in mind when interpreting any differences in outcomes between the two groups, as the participants may have entered the programme with a marginally higher prior attainment profile.

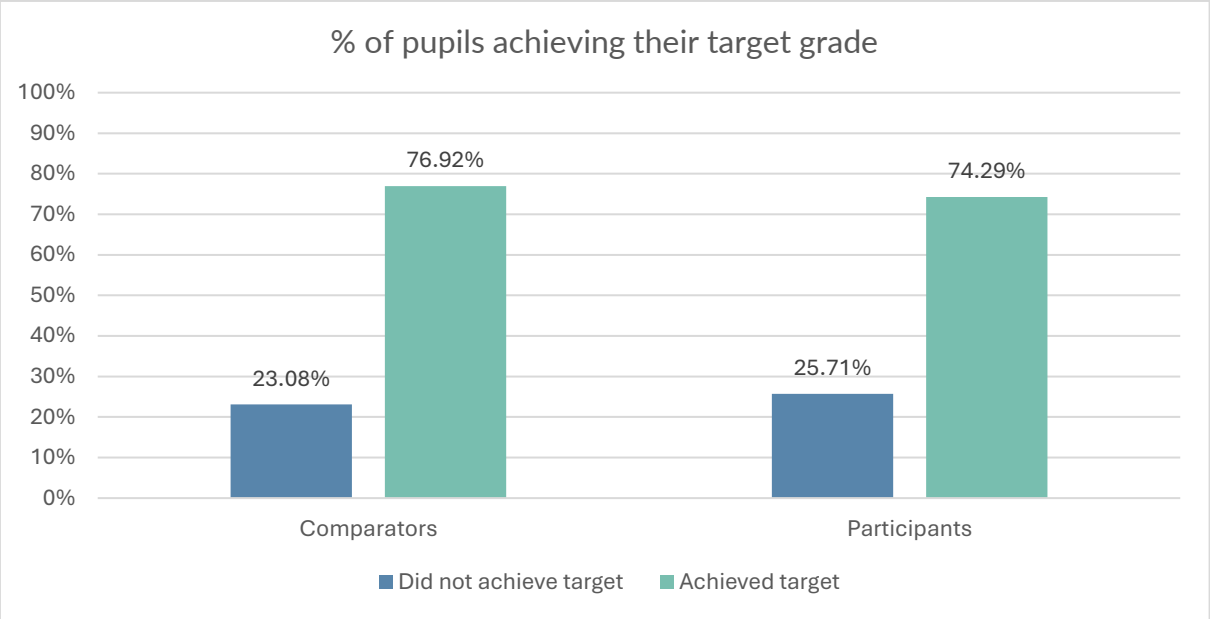

Figure 1 shows the percentage of participants and comparators that achieved their target grades

When looking at the percentage of students that went up a whole grade from Spring Mocks to GCSE grades it was found that 15% of the 35 participants achieved this compared to 13% of the 26 comparators. In addition, 18% of the 35 participants went up a whole grade from their target grades to their final GCSE grades while only 10% of comparators went up a whole grade. Participants showed slightly better grade progression than comparators, with the difference most notable when comparing final GCSE grades to target grades, participants were roughly twice as likely to exceed their target grade (18% versus 10%). However, the small sample sizes mean these findings should be treated with caution.

Attainment- Ernest Bevin

Key finding: Pupils from Ernest Bevin showed a larger increase (13.6 percentage points) than their comparator peers (9.38 percentage points) from their Mock grade in the Autumn term to their final GCSE grade in September 2025. This difference was not statistically significant (*p* = 0.055, n = 32).

Ernest Bevin school shared Autumn term mock grades which enabled a comparison of their Mock GCSE grades from the start of the year to their final GCSE grades at the end of the academic year. This showed an increase from an average grade of 3.24 in the Autumn term to 4.32 by the end of the academic year. The comparator group also showed an increase in attainment from 2.79 to

3.54. The absence of statistical significance may be attributed to insufficient statistical power resulting from the small sample size. Despite this, the improvement is particularly noteworthy given that participants started with higher baseline grades than their peers in the comparison group. This is known as the ceiling effect: pupils with higher initial scores typically have less room for improvement compared to those starting at lower levels.

Table 61 shows the attainment findings for Autumn to Summer term for Ernest Bevin

Attainment	Type of Pupils	Sample size	Baseline – Autumn Mock	Endline - GCSE	Percentage point difference	Statistical significance
Science	Comparators	15	2.79	3.54	9.38	$p = 0.055$
	Participants	17	3.24	4.32	13.60	

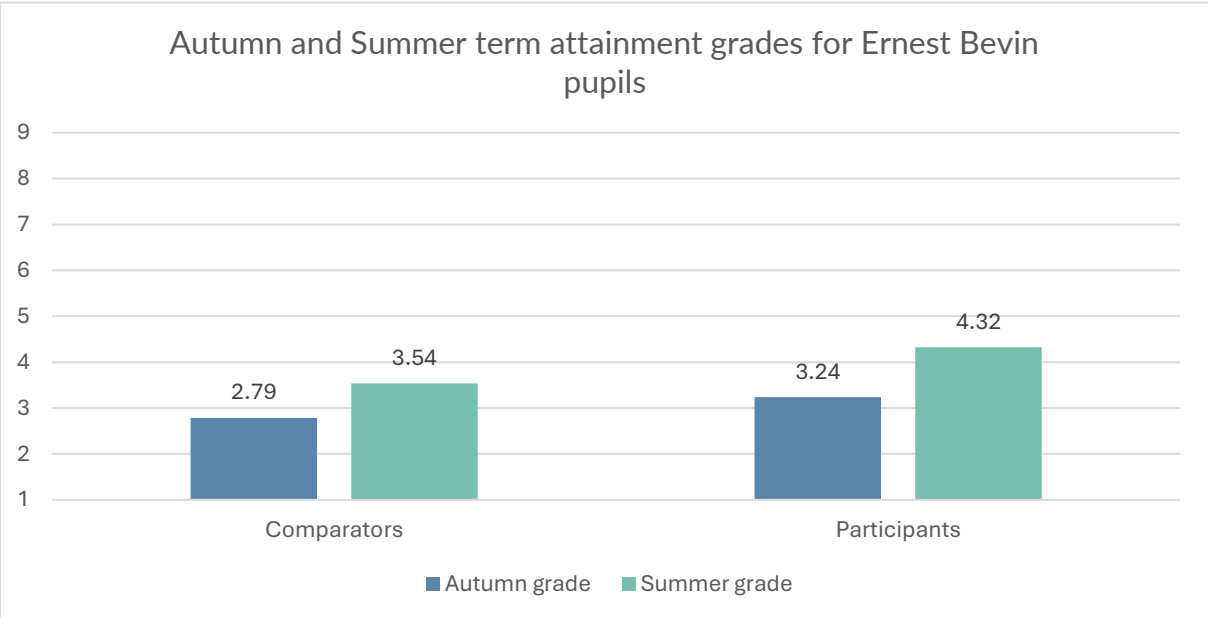

Figure 2 shows the change in attainment grades from Autumn to Summer term for Ernest Bevin pupils

Looking closer at the change in grades from the Spring mock grade to the final GCSE grade in September 2025 for Ernest Bevin pupils, there was a smaller difference between participants and comparators compared to the scores overall. This difference was not statistically significant suggesting the difference could be due to chance.

Table 72 shows the change in grades from Spring to Summer term for Ernest Bevin pupils

Attainment	Type of Pupils	Sample size	Spring mock grade	Endline - GCSE	Percentage point difference	Statistical significance
Science	Comparators	15	2.63	3.54	10.71	P=0.46
	Participants	17	3.65	4.32	8.46	

Key finding: 94% of participants at Ernest Bevin achieved their target grade compared to 100% of their comparator peers. This was sixteen out of the seventeen participants that achieved their target grade. This difference was not statistically significant (*p* = 0.36, n = 32)

Target grade attainment was very high among both Ernest Bevin participants (94%) and their comparator peers (100%). While comparators achieved universal success in meeting their target grades compared to 94% of participants, this percentage point difference was not statistically significant. The modest sample size (n = 32) provides limited statistical power to detect small differences.

When looking at the range of target grades for the 17 participants at Ernest Bevin, this was 3.5, with a mode of 3 and 4 and a median of 3.5. For the comparator group of 15 students the range was 2, the mode was 2.5 and the median was 3. The participant group had higher target grades on average (median 3.5 vs 3) and greater spread (range 3.5 vs 2) than the comparator group, suggesting the two groups in this school were not as well-matched on prior attainment. This is an important limitation to note when interpreting any outcome comparisons.

The percentage of students that went up a whole grade from Spring Mocks to GCSE grades was 11% of the 17 participants compared to 8% of the 15 comparators. In addition, 13% of the 17 participants went up a whole grade from their target grades to their final GCSE grades while only 5% of the 15 comparators went up a whole grade. This suggests participants at Ernest Bevin showed slightly better grade progression than comparators, with the difference most pronounced when comparing final GCSE grades to target grades. Participants were more than twice as likely to exceed their target grade (13% versus 5%). As with the overall findings, the very small sample sizes mean these results should be interpreted cautiously.

Attainment- Burntwood

Burntwood shared Spring mock GCSE grades and final GCSE grades in September 2025. Findings showed a negative change in Science GCSE attainment in their final GCSE grades compared to their mocks (-4.51 percentage points) while the comparator group showed a positive change (+4.55 percentage points). This change was statistically significant suggesting the change is less likely to be due to chance

Table 83 shows the change in grades from Spring to Summer term for Burntwood pupils

Attainment	Type of Pupils	Sample size	Spring grade	Endline - GCSE	Percentage point difference	Statistical significance
Science	Comparators	11	3.95	4.32	4.55	P=0.026
	Participants	18	4.92	4.56	-4.51	

Key finding: A greater percentage of Burntwood participants (56%) achieved their target grade in comparison to their comparator peers (45%). This was not statistically significant (p = 0.61, n = 29).

Compared to Ernest Bevin, fewer participants and comparators achieved their target grades. However, this positive comparison between participants and their peers could indicate that Science Stars is having a positive impact on their ability to reach the standard expected from them. Yet, the combination of a small sample size and moderate success rates in both groups limits our ability to draw firm conclusions.

The range of score for Burntwood's 18 participants was 3, with a mode of 4 and a median of 4.5 while the 11 comparators had a range of 3, a mode of 4 and a median of 4. This suggests that the two groups had similar target grade profiles, with both groups most commonly targeted at a grade 4. However, Burntwood's participants were slightly higher attaining on average, with a median of 4.5 compared to 4 for the comparators.

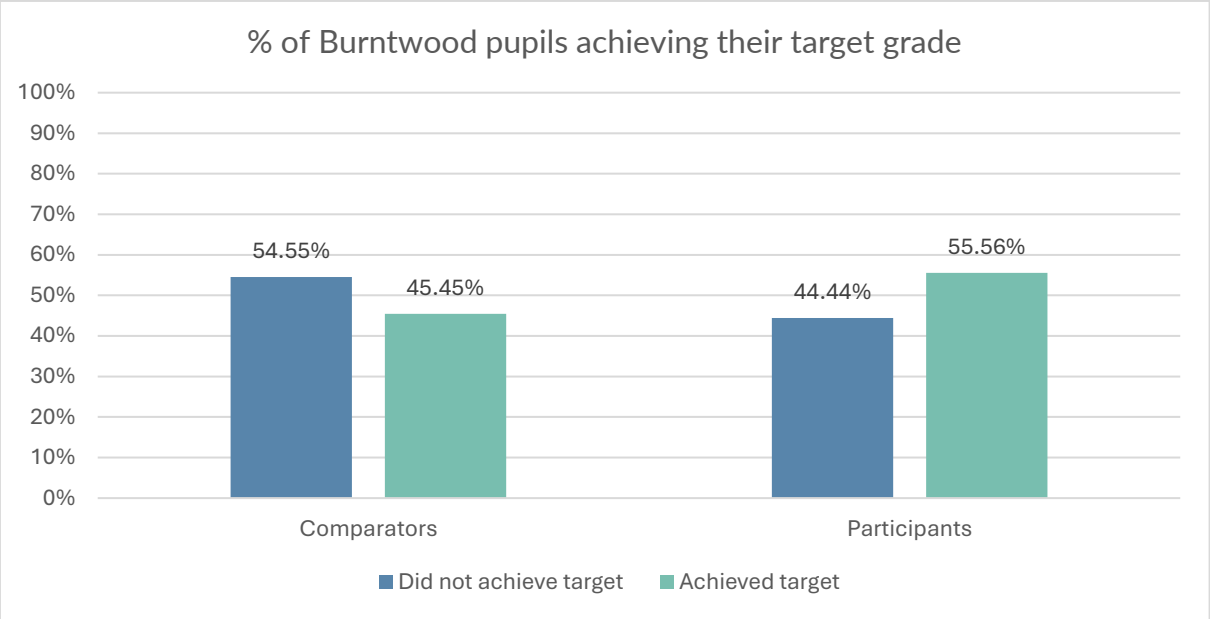

Figure 3 shows the percentage of participating and comparator pupils that met their target grade from Burntwood

The percentage of students that went up a whole grade from Spring Mocks to GCSE grades was 3% of the 18 participants compared to 5% of the 11 comparators. In addition, 5% of the 18 participants went up a whole grade from their target grades to their final GCSE grades while only 5% of the 11 comparators went up a whole grade. For this group, participants showed no clear advantage over comparators, progression rates from Spring Mocks to GCSE were slightly lower for participants than comparators (3% versus 5%), and the proportion exceeding their target grade was identical for both groups (5%).

Cross- year: Non-cognitive and SEMH skills- All pupils

Key finding: This year saw a continuation of last year's trend; participants saw positive progress in metacognition, test anxiety and self-efficacy. The positive progress in self-efficacy was larger than the previous year, while the progress in metacognition and test-anxiety was smaller.

There was a noticeable increase in self-efficacy scores compared to last year's cohort. This is extremely positive as this finding was also statistically significant when compared to the outcomes of this year's comparator group, suggesting the programme successfully improved participants' self-efficacy rather than this improvement being due to chance. Although participants' positive

progress is clear, it is worth investigating why the rate of progress has slowed down from last year, particularly for metacognition and test anxiety. Some further exploration of the implementation of the tutoring in both of the schools could give some context for this lower impact.

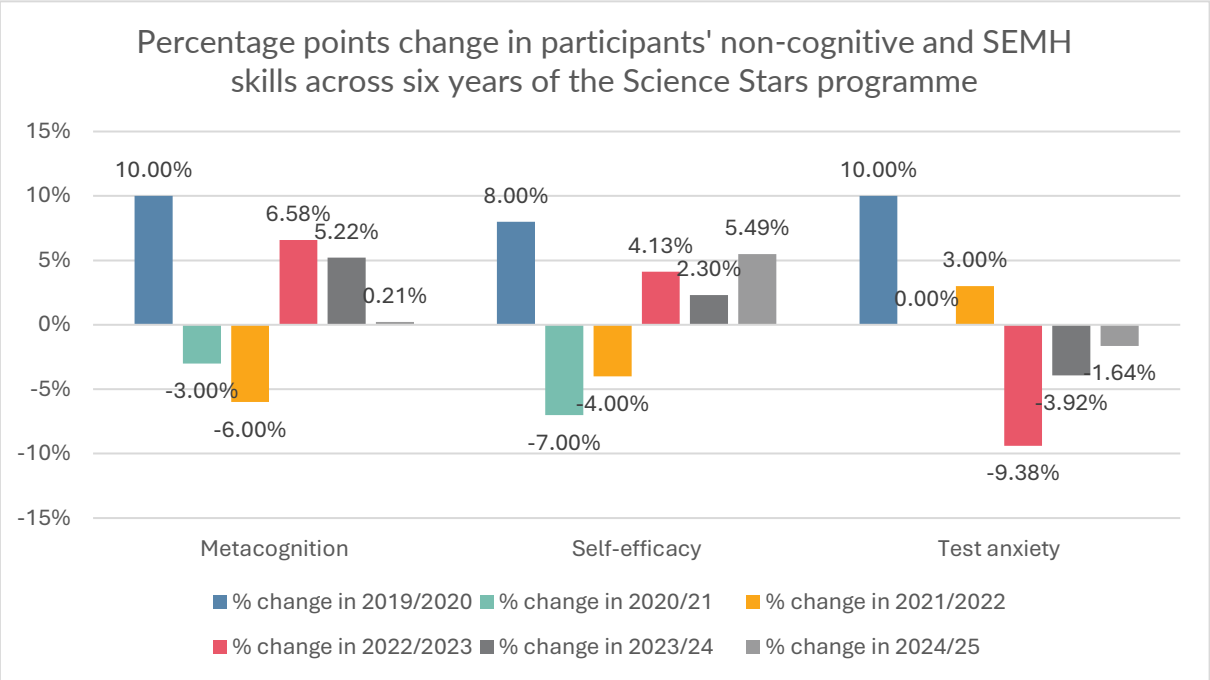

Figure 4 shows the change in participants non-cognitive and SEMH skills over the last six years

Cross- year: Attainment- All pupils

A comparison of changes in science grades between this year and previous cohorts was not possible due to differences in data availability. Previous years had comprehensive grade data for the entire cohort across all terms, whereas this year only target grades, spring mock grades and final GCSE grades in September 2025 were available for the full cohort.

Key finding: The percentage of participants achieving their target grade decreased by 8.57% from last year to this year.

Despite a decrease since the previous year, the percentage still shows that the majority of pupils (74%) managed to meet their target grade. This change could be influenced by the differences between schools in how they set their target grades in the Spring term.

Table 94 shows the percentage of pupils that met their target grade each year

Attainment	Type of pupil	% of pupils 2020/2021	% of pupils 2021/2022	% of pupils 2022/2023	% of pupils 2023/2024	% of pupils 2024/2025
Science	Comparators	33%	61%	42.42%	66.67%	76.92%
	Participants	58%	64%	60.61%	82.86%	74.29%

Key finding: A smaller proportion of participants achieved their target grade compared to comparators. This difference between groups did not reflect the positive difference from previous years.

Previously, more participating pupils achieved their target grade than comparator pupils. Whereas this year, the was a small negative difference between participants and comparators achieving their target grades. This suggests, further consistency in target setting may need to be considered across schools to make a more robust comparison.

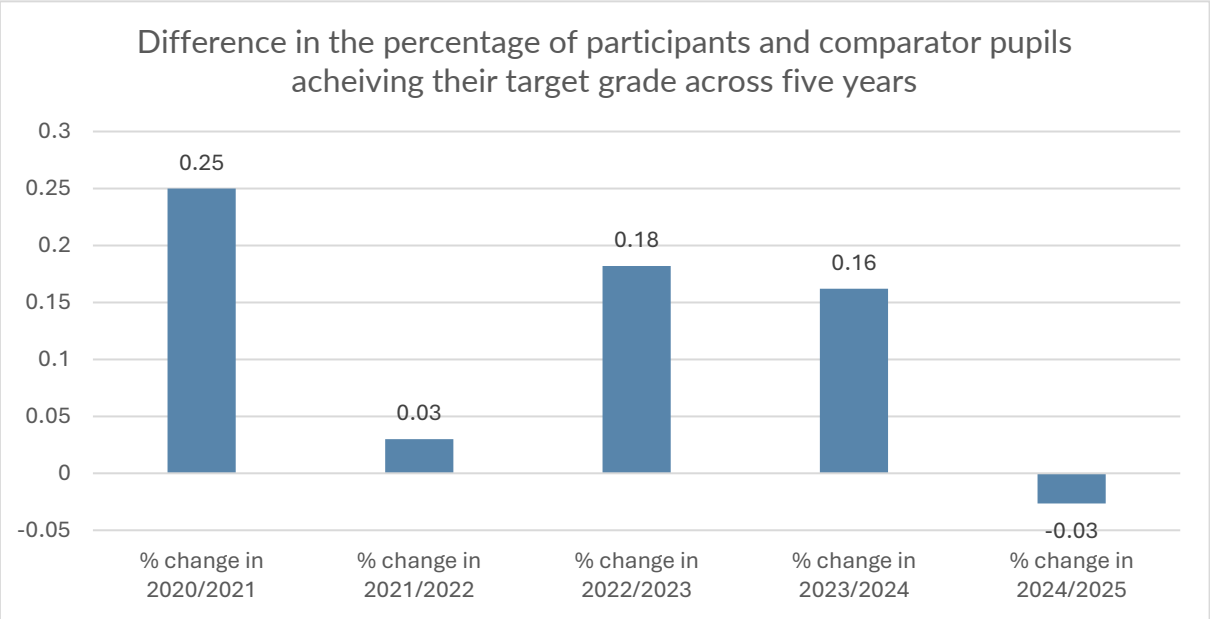

Figure 5 shows the difference in the percentage of participants and comparator pupils achieving their target grade across five years.

Programme Delivery

The qualitative focus groups and interviews explored feedback from both student ambassadors who delivered the tutoring sessions and a teacher from each school. This enabled some exploration of successes of the programme as well as areas for improvement.

Student Ambassadors' motivation

Tutors at Ernest Bevin were motivated to become student ambassadors because they wanted to gain experience in teaching and help other pupils who may need a different approach to learning. In addition, in the focus group, they spoke about their own experiences and how this motivated them to provide help for others who need it.

Tutors at Burntwood were motivated to become student ambassadors because they liked what the programme was aiming to achieve:

"...helping students from different backgrounds perform at their best."

- *Student ambassadors, Burntwood*

They also reflected on their own experiences and were motivated to give something back to the community.

Successes and Areas of Improvement of the Science Stars programme

Student Ambassadors

Student Ambassadors reflected on the benefits of small groups and how this enabled a good level of individual communication and group work. They enjoyed the structure and frequency of the sessions and thought that this consistency was appreciated by the pupils too. However, it was noted by student ambassadors from Burntwood that the timing of the sessions towards the end of the term impacted pupils' motivation as they seemed to get progressively more tired and therefore less motivated.

Student ambassadors found it beneficial to speak to more experienced tutors and receive tips and techniques in the training on how to deliver the topics. One example they mentioned was how to

ask questions to the pupils effectively. They also found it more engaging with group activities and opportunities to align with other tutors on different activities. The communication via the group chat was also deemed useful.

“The support is quite good in terms of what we need to help and deliver to the students.”

- *Student Ambassador*

The student ambassadors reflected on the structure of the tutoring sessions and thought that the quizzes at the end worked well and aided their planning and preparation for the next session:

"You can kind of prepare sessions a lot quicker because you already know the information…this has definitely made it possible for me to balance this commitment with my studies. If it hadn't been the case, I don't think I could have done it otherwise."

- *Student Ambassador*

Student Ambassadors recommended the following as elements to improve for the following year’s programme:

- Further direction on the content of what the pupils need to know. One suggestion was to make the slides more representative of the depth to go into. Another was to have information on pupils’ academic abilities and needs before the session, so that the questions and lessons could be more targeted to their needs.
- To have longer sessions of 1.5 hours to give more time to cover all of the content in detail.
- Further information on accessing the school and the availability of resources would be beneficial for future lesson preparation.

Attendance seemed to be one of the most profound barriers to delivery. Student ambassadors made some suggestions to improve this below:

- Some contextual information on the pupils’ other commitments so that if they were unavailable for the session due to other commitments/ activities, the space in the session could be taken by others.
- Communication to staff in the schools on when and where the sessions are taking place to inhibit detentions from being issued at the same time.

- One student ambassador mentioned that they turned up twice to no pupils, which was a waste of their time- pupils could be held more accountable if they miss sessions and explain to the teacher why the sessions were missed.
- A session at the start could take place with parents to address pupils' commitment to attend, or parents could have access to the schedule for Science Stars.
- A report card could be issued halfway through the sessions that could be sent to parents stating the pupils' attendance.

Teachers

One teacher expressed the benefit of having people from the community helping and supporting others and the positive link this creates with the university. In addition, the positive relationships with the student ambassadors was mentioned:

"...they come in and they have this real care with the kids. I can't count the number of times kids have shown me revision mats or knowledge organisers or question papers that I didn't print and I'm thinking they've printed stuff outside of the school, brought it in for the students, and you know that little act of students seeing people do extra work for them, makes them want to do extra work"

- *Teacher*

Both teachers acknowledged that their previous experience and understanding of the processes and programme were useful in informing the planning and preparation.

Both teachers commented upon the attendance of the pupils. One teacher gave an example of a pupil whose attendance isn't normally very good, but who would make the effort to come in every week for something that was helping them. It was noted that attendance dwindled towards the end of term as expected, but it was an improvement over the previous year.

One teacher remarked that there was a good routine for communication but suggested that weekly feedback could be given earlier so that they are able to report back to the pupils ahead of the next session where needed.

Conclusion and Recommendations

Overview of survey and attainment data

This year, survey data showed positive change for some of the non-cognitive skills and negligible change for others. There was limited change (0.21 percentage points) in revision skills from baseline to endline for participating pupils, suggesting a very small change in pupils' understanding of their strengths and weaknesses. Yet, for the comparator group, there was a negative change of -6.63 percentage points. This suggests that the Science Stars programme may have supported pupils' revision skills to an extent, preventing a decrease in their metacognition and revision skills. Participating pupils' confidence in science improved (+5.49 percentage points), whereas the comparator group showed a negative change (-19.79 percentage points). This positive change for participating pupils could highlight the impact of the Science Stars programme upon pupils' self-efficacy compared to those who did not participate. Participating pupils' test anxiety decreased (-1.64 percentage points), suggesting a slight improvement in their feelings towards tests and exams. Whereas comparator pupils showed an increase in test anxiety (5.03 percentage points). This suggests that the programme overall had a positive impact on pupils' non-cognitive and SEMH outcomes as intended. Yet, when this is broken down by school, some differences come to light. Ernest Bevin school seemed to follow the positive trend for all non-cognitive and SEMH outcomes, whereas findings for Burntwood were mixed. Consideration of a larger sample and further investigation into the implementation of the programme within the two schools could be explored further.

In summary, overall, all of the participants' improvements in non-cognitive skills were better than those of their comparator peers. In addition, metacognition and self-efficacy findings were statistically significant ($p=0.02$ and $p=0.01$, respectively). This suggests that the limited change in revision skills was unlikely due to chance, suggesting the Science Stars programme had a genuine effect on pupils' metacognition, revision skills and their understanding of their strengths and weaknesses. Despite this change being relatively small, in comparison to their comparator peers, there was no negative change. In addition, the statistically significant improvement in self-efficacy indicates a genuine positive impact of the Science Stars programme.

On average, participants saw a small increase in their Science grade from their Spring mock grade to their final GCSE grade (+1.79 percentage points), and a large proportion of participants achieved their target grade (74%), indicating some enhanced understanding of Science GCSE content over the course of the programme. Similarly, comparator peers showed further increase from their Spring mock grade to their final GCSE grade with a 7.93 percentage point change and a slightly larger percentage meeting their target grades (76.92%). However, none of the differences between participants' progress and their comparator peers were statistically significant for academic attainment. This means that the observed differences could simply be due to chance rather than a genuine effect of the programme. Findings were different between the two schools, suggesting that some consistency in setting target grades could be considered or further collaboration between the student ambassadors and teachers who are implementing the programme could be explored. Some of the qualitative findings also looked at tailoring sessions more towards the pupils' needs, so further preparation on pupils' academic attainment and needs prior to the sessions starting could be considered. While there are differences in pupils' learning styles and behaviours, it is also important to consider that the student ambassadors may have different approaches to tutoring, and so some further consideration of these different approaches may explain some of the differences in findings between the two schools.

Evaluation Recommendations

- Investigate external factors that may have affected pupils' lower gains in metacognition and test-anxiety compared to the previous year. This could involve further exploration of tutoring strategies and ability to tailor learning to the pupils' needs.
- Consider a more rigorous matching procedure such as propensity score matching so more robust conclusions can be drawn between participating and comparator pupils.
- Consider exploring pupil voice where possible to triangulate findings.
- Continue to increase sample size where possible. An increased sample size will improve the robustness of the findings. It could be worth exploring whether the programme could be scaled and retain the same benefits with a larger group.

Programme and Delivery Recommendations

- Consider sharing further details on pupils' academic abilities and needs with the student ambassadors prior to the sessions starting.
- Consider strategies to monitor the attendance of the pupils.
- Consider an exploration of teaching and learning styles prior to the programme starting so that sessions can be tailored to meet individual needs.

- Ensure feedback to schools is always timely so that those involved can be addressed prior to the next session.
- Make sure all data collection is consistent across schools so that comprehensive data is shared to strengthen findings.
- We recommend investigating what Ernest Bevin was doing particularly well to lead to such positive results, and exploring how this could be emulated across other schools.

Glossary

Evaluation terminology

Academic attainment

This refers to test scores in academic subjects such as maths, science, English etc. Some evaluations will compare pupils' attainment in tests for these subjects at the start (baseline) and end (final) of an evaluation to see whether they have made progress over time.

Academically validated measures

These are scales to measure social and emotional skills linked to academic achievement and long-term life outcomes that have been developed and peer reviewed by academic researchers within the fields of education and psychology. These have been developed to ensure:

- Predictive validity. These skills have been shown to be closely related to desirable life outcomes such as educational achievement, employability and earnings potential, or long-term health and life satisfaction. (In psychometrics, predictive validity is the extent to which a score on a scale or test predicts scores on some criterion measure. For example, the validity of a cognitive test for job performance is the correlation between test scores and, say, supervisor performance ratings.)
- Construct validity. The measure tests for the skill that it says it does, as defined in the literature.
- Test-retest validity. The results stay the same when tests are repeated.

Baseline

The initial assessment of pupils' attainment or social and emotional skills, at the start of an evaluation.

Change over time

The difference between a pupil's baseline result and their final result, either for attainment or social and emotional skills. This indicates progress made during participation in the programme. This will begin to indicate whether the programme has had an impact on pupils, though we must also

account for other factors that could lead to this change, which is why we recommend the use of control groups and qualitative analysis.

Control Group

A control group is composed of students who do not participate in the programme and who closely resemble the pupils who take part in the programme in attainment and demographic traits. It is used to get an indication of whether a change in results over the course of the programme can likely be attributable to the programme itself, or whether results were likely to change over time in any case. Also known as a comparison group.

Evaluation

An evaluation is set up to measure the impact of a particular programme. This will involve monitoring the programme over a specified period, for one or more groups, in order to evaluate the progress participating pupils make. One programme can involve multiple evaluations, and we recommend gathering data across multiple time points to ensure valid and reliable results are generated.

Evaluation Group(s)

An evaluation will either cover one specific group of pupils, who all participate in the programme (e.g. a new programme trialled in one class, or an intervention with one small group). Or, the evaluation may cover multiple evaluation groups (e.g. as several small-group interventions, or with multiple classes carrying out the same programme). In the case of multiple evaluation groups, it can be useful to compare the outcomes for different groups to build up a stronger data set, as well as to compare differences in implementation to see whether this has an effect on results.

Final

The final assessment of pupils' attainment or social and emotional skills at the end of an evaluation.

Matched Pupils

Matched Pupils are pupils who carried out both a baseline and a final assessment at the start and end of the evaluation. It can be useful to consider results from Matched Pupils only because this means only including those pupils who participated in the full duration of the programme.

Outcomes

We use outcomes to refer collectively to any social and emotional skills and academic attainment scores that are being measured over the course of an evaluation.

Participating pupils

The group of pupils participating in the evaluation, and not forming part of a control group.

Programme

This could be any intervention, project or programme run in school with the aim of improving pupil outcomes or life chances. ImpactEd works with schools to build evaluations of their programmes in order to better understand whether they are having their intended impact.

Skills measures

We use a set of academically validated skills measures to assess pupils' social and emotional skills. See Our Metrics, below, for details of each measure we use.

Social and emotional skills

The term 'social and emotional skills' refers to a set of attitudes, behaviours, and strategies that are thought to underpin success in school and at work, such as motivation, perseverance, and self-control. They are usually contrasted with the 'hard skills' of cognitive ability in areas such as literacy and numeracy, which are measured by academic tests. There are various ways of referring to this set of skills, such as: non-cognitive skills, twentieth century skills and soft skills. Each term has pros and cons; we use social and emotional skills for consistency but we recognise that it does not perfectly encapsulate each of the skills that come under this umbrella.

Statistical analysis terminology

Statistically significant

A result has statistical significance when it is very unlikely to have occurred given the null hypothesis. In other words, if a result is statistically significant, it is unlikely to have occurred due purely to chance.

P Value

A p-value is a measure of the probability that an observed result could have occurred by chance alone. The lower the p-value, the greater the statistical significance of the observed difference. Typically a p-value of ≤ 0.05 indicates that the change was statistically significant. A p-value higher than 0.05 (> 0.05) is not statistically significant and indicates strong evidence for the null hypothesis; i.e. that we cannot be confident that this change did not occur due purely to chance.

Measures for social and emotional skills

Metacognition

Metacognition means 'thinking about thinking': pupils' ability to think explicitly about their own learning (Flavell, 1979; Higgins et al., 2016). It is strongly associated with academic progress and improves other skills required for learning, such as critical thinking. Metacognition enables pupils to develop strategies to plan, monitor, and evaluate their learning.

Self-efficacy

Self-efficacy is a measure of pupils' belief in their ability to achieve a specific task in the future. Self-efficacy is correlated with higher academic achievement and persistence, and also contributes to pupil wellbeing. (Gutman & Schoon 2013, DeWitz et. al. 2009).

Test anxiety

Test anxiety is concerned with pupils' emotional responses to tests (Pintrich and De Groot, 1990). Greater levels of test anxiety can result in worse performance in exams, but may in some situations be linked to increased motivation and self-regulation.