ImpactEd Evaluation

Science Stars

23/24 IMPACT REPORT

www.evaluation.impactedgroup.uk

Contents

Executive Summary

Science Stars is a tutoring intervention delivered by St. George's, University of London. It aims to improve the science GCSE attainment of Year 11 pupils. 23/24 was **the fifth year of the Science Stars tutoring programme and this evaluation has found potential benefits** but where **further investigation** with a more **rigorous control** would help establishing **further positive impact.**

This year, the programme was delivered in-person at two schools. One of these schools was a school who had not participated in the programme before, Burntwood. This was a shift from last year, where one school delivered the programme in-person, and the other school delivered it online. From key findings last year that showed that pupils were making less progress across non-cognitive and social, emotional and mental health skills when the programme was delivered online, ImpactEd Evaluation recommended in-person delivery for both schools. This was implemented for 23/24.

This evaluation adopted a mixed-methods approach and used quantitative attainment and non-cognitive surveys as well as qualitative interviews and focus groups. Whilst the first two years of evaluation (2019/20 and 2020/21) showed positive impacts of the programme, the next year (2021/22) showed the beginning of some negative trends. However, the evaluation of last year's programme (2022/23) showed positive results, and this year (2023/24) continues the positive trend.

Although the positive trends from last year have continued, the impact appears to be less pronounced this year. Overall, analysis of academic outcomes showed that **participating pupils improved their non-cognitive and SEMH skills, as well as their science GCSE grades, more than their peers** in the comparison group.

Key findings

The 2023/2024 findings present a nuanced picture of pupil performance across two schools. While participants generally demonstrated improvements compared to their peers, the outcomes varied significantly between Ernest Bevin and Burntwood. Ernest Bevin pupils showed consistently stronger progress across all measured indicators, substantially outperforming their comparator peers. In contrast, Burntwood's participants experienced more mixed results, with their comparator peers often performing slightly better. It should be noted that the creation of the comparator group involved identifying clusters of pupils with similar academic profiles so comparisons should be drawn judiciously as this approach may not account for all relevant underlying factors.

Despite the comparative performance data, qualitative insights from Burntwood reveal an encouraging narrative of the programme's effect on individual pupils, suggesting that statistical averages do not capture the full complexity of educational interventions and pupil development.

23/24 - Findings

1. On average, participants' metacognition increased (+5.22 percentage points) slightly more than their comparator peers (+2.66 percentage points), suggesting the Science Stars programme has positively impacted participants' revisions skills and understanding of their own learning.
2. On average, participants improved their self-efficacy (+2.33 percentage points) more than their comparators peers (-0.07 percentage points), suggesting Science Stars had a positive impact on participants' confidence in science.
3. On average, participants decreased their levels of test anxiety (-3.92 percentage points) more than their comparator peers (-1.43 percentage points), suggesting that Science Stars had a positive impact on participants' anxiety around tests and exams.
4. The difference between participants and comparator peers across all three measures at Ernest Bevin is more pronounced than the overall trend, but the differences are not statistically significant.
5. 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 a random change rather than a true effect.
6. The progress made in GCSE science was more positively pronounced on participants in Ernest Bevin than in the overall picture. On average, their grades increased from 3/53 to 4.47 (+0.95), whereas their comparator peers' grade increased less (3.18 to 3.38 (+0.21 grades); this difference was statistically significant ($p < 0.05$, n = 36).
7. On average, participants increased their science grade from a 4.14 to a 4.69 (+0.54) which was more than the progress made by comparators (3.83 to 4.12, +0.29). This difference was not statistically significant ($p = 0.13$, n = 68).
8. 83% of participants achieved their target grade which was more than the 67% of comparator peers who reached their target. This difference between participants and comparators was not statistically significant ($p = 0.13$, n = 68).
9. The absence of statistical significance in the two findings above may be attributed to insufficient statistical power resulting from the small sample size. The differences above may also be better explained by random chance rather than there being a genuine difference between the two populations.

Longitudinal - Findings

1. 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 all these three measures this year, however, was smaller than the progress made last year.

2 2023/24 saw a larger percentage point increase in Science grades than in 2022/23 and reached similar levels to previous years of the Science Stars programme.

3 The difference between participants' and comparator peers' percentage point progress made from their Autumn term mock to their final GCSE grade decreased from 2022/23 to 2023/24.

4 The positive difference in the percentage of pupils achieving their target grade between participating pupils and their comparator peers has slightly decreased from 2022/23 to 2023/24.

Recommendations

Following another successful year, we propose recommendations for programme evaluation, pupil-focused initiatives and broader programme delivery:

- Investigate the differential impact on the two schools' pupils' non-cognitive and SEMH skills.
- Investigate differential impact on academic progress.
- Analyse external factors that may have affected pupils' lower gains in SEMH and non-cognitive skills compared to the previous year.
- In 2024/25, Year 11 pupils' target grades will be constructed differently to previous years due to covid-19 having impacted their Year 6 SATs in 2019/20; St George's and ImpactEd Evaluation to discuss how this may impact future evaluations.
- Consider a more rigorous matching procedures so more robust conclusions can be drawn between participating and comparator pupils.
- Consider incorporating additional stress and anxiety management workshops.
- Teachers and tutors could share best practices among their peer groups to maximize programme impact.
- Tailor teaching support, such as grouping students by academic ability and integrating diverse learning skills into teaching methods.
- Enhanced communication processes, including better coordination with school staff and alignment with school schedules.
- Conduct a review of training materials to ensure that training is relevant and implementable by student teachers.
- Ensure that all trainings are attended by all student teachers.
- Implement a system for early planning of programme to ensure schools feel confident establishing the programme.
- Acknowledge and address the logistical and safety concerns, particularly for students traveling at night.

1. Introduction

St George's, University of London, is an independent university dedicated to medical and health science education, training and research, affiliated with the University of London. With a strong historical commitment to widening participation activities, St George's is now increasingly working across the whole student lifecycle to support students from under-represented backgrounds. This year, St George's has run the Science Stars programme for a fifth 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 sustained tutoring intervention designed to support Year 11 students to prepare for GCSEs and ultimately increase their attainment in science. The programme is delivered in-person by current students at St George's, University of London – 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 students in Year 11. The key aims and objectives of the programme for participating students 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

2. Methodology

The methodology section consists of key research questions, outcome measures, the evaluation design for data collection, and limitations of the approach.

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

Outcome	Quantitative Measure	Qualitative Measure
Improved revision skills	MSLQ Metacognition	Focus groups with graduate tutors and interviews with teachers (all outcomes)
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 fifth annual evaluation of this programme and was conducted in 2023/24. All the data was collected between Autumn Term 2023 and Autumn Term 2024.

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 matched 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. This group will be referred to as comparator group throughout the report. It should be noted 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.

Although there are some limitations of 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:

- **Attainment data** was used to evaluate the impact of the programme on pupil's academic progress,
- **Pupil survey data** was used to evaluate the impact of the programme on pupils' non-cognitive outcomes,
- **Qualitative research** and **delivery data** was used to evaluate the success of the implementation of the programme.

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.

Table 2

Data	When?	Which pupils?	Matched Sample Size	
			Ernest Bevin	**Burntwood**
Autumn Mock exam	Autumn Term 2023	Participating	19	16
		Comparator	17	16
Final GCSE results	September 2024	Participating	19	16
		Comparator	17	16

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 with associated 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. The evaluation followed a pre-post-test design. 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 3

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 survey will be supplemented by qualitative data that has been drawn out by the four focus groups with eight Science Stars tutors with and two one-to-one interviews with the group assistants, one from Burntwood and the other from Ernest Bevin.

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

Table 4

Data	When?	Which pupils?	Matched Sample Size	
			Ernest Bevin	**Burntwood**
Meta-cognition baseline	Autumn Term 2023	Participating	8	18
		Comparator	15	17
Meta-cognition endline	Spring Term 2024	Participating	8	18
		Comparator	15	17
Self-efficacy baseline	Autumn Term 2023	Participating	12	17
		Comparator	12	14
Self-efficacy endline	Spring Term 2024	Participating	12	17
		Comparator	12	14

Test anxiety pre-survey	Autumn Term 2023	Participating	14	20
		Comparator	17	18
Test anxiety post-survey	Spring Term 2024	Participating	14	20
		Comparator	17	18

Qualitative Research: Design, Sample and Analysis

Focus groups were conducted with tutors leading sessions at both schools. Four tutors participated in focus groups for Ernest Bevin, and four tutors participated in focus groups for Burntwood. 1:1 interviews were conducted with the relevant schoolteacher 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.

National Benchmarks for non-cognitive outcomes

Benchmarks used for non-cognitive comparisons use data from the School Impact Platform.

Benchmarks were available for metacognition and self-efficacy, but not for test anxiety. These two benchmarks were constructed in two slightly different ways because the raw data available varied between the measures.

Metacognition Benchmark Construction:

- Baseline: Average of pupil responses from Autumn Term 1 and 2 (October-November 2023/24)
- Endline: Average of pupil responses from Summer Term 1 (2023/24)
- Calculation: Difference between baseline and endline divided by 6 to produce a percentage point difference

Self-Efficacy Benchmark Construction:

- Baseline: Average of secondary school pupils' survey responses during the same period as Science Stars participants' baseline
- Endline: Average of pupils' survey responses during the same period as Science Stars participants' endline
- Calculation: Difference between baseline and endline divided by 6 to produce a percentage point difference

The methodology ensures a comparative analysis by matching time periods and calculating percentage point changes for both benchmarks and participants.

Table 5

Measure	n for baseline	n for endline

Metacognition	9050	1340
Self-efficacy	828	126

Analysis Terminology: Percentage vs Percentage Points

Throughout the report, the terms percentage vs percentage points will be used. Please note the difference between the meanings in the definitions below:

- **Percentage change** quantifies the change we observed as a proportion of the value we started from.
- **Percentage point change**, on the other hand, quantifies the change we observed in absolute terms (i.e. not relative to the starting point). For example, if 50% of pupils answer 'yes' to a certain question in our baseline survey, but then, later on, 55% of pupils answer 'yes' to that same question in our endline survey, this is a change of **5 percentage points** but a change of 10% (since the difference, 5, is 10% of the starting value, 50).

Limitations

There some limitations of this evaluation design worth noting:

- As the comparison group was not randomised, there may be unobservable characteristics affecting performance beyond prior attainment.
- Particularly when looking at 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.
- This report examines performance differences between two schools: one with multiple years of programme experience and another in its first year of implementation.
- School by school breakdown was not conducted for longitudinal analysis because one of the two schools had only engaged with the programme for one year.
- Most of the time, qualitative analysis is used to qualify and to explain differential impact on the two schools participating in the programme.
- Spring mock data was not included in the analysis this year.

3. 23/24 - Outcomes

This year saw Science Stars participants increase all three of their non-cognitive skills, implying that they improved their revisions skills (+5.22 percentage points), increased their confidence in science (+2.3 percentage points), became less anxious about tests and exams (-

3.92 percentage points), and now have a better understand of their own learning strengths and weaknesses (+5.22 percentage points). On average, participants also saw an increase in their Science grade from their Autumn Mock to the final GCSE grade (+15 percentage points), and a large proportion of participants achieved their target grade (83%), indicating their increased ability to answer exam questions, understand Science GCSE content and increase in their academic attainment.

All the participants' improvement in non-cognitive skills and their academic attainment were better than their comparator peers but none of the differences between participants' progress and their comparator peers were statistically significantly. This means these changes may be better explained by sampling rather than genuine difference in groups. While the overall trend appears positive, the data for Burntwood reveals nuanced variations. Specifically, some measures showed decreases in participants, and in some instances, comparator pupils demonstrated higher performance than participants.

Non-cognitive and SEMH skills – all pupils

Key finding: On average, participants' metacognition increased (+5.22 percentage points) slightly more than their comparator peers (+2.66 percentage points), suggesting the Science Stars programme has positively impacted participants' revisions skills and understanding of their own learning.

The **difference between participants and comparator pupils was not statistically significant**. Participants also saw a greater percentage point increase than the national average (+0.64 percentage points).

Table 6

	Type of Pupils	Sample size	Baseline	Endline	Difference	Percentage point difference	Statistical significance
Metacognition	Comparators	32	4.41	4.57	+0.16	2.66 %	$p = 0.57$
	Participants	26	4.22	4.53	+0.31	5.22%	
	Benchmark	Refer to methodology	3.55	3.59	+0.04	0.64%	

Key finding: On average, participants improved their self-efficacy (+2.33 percentage points) more so than their comparators peers (-0.07 percentage points), suggesting Science Stars had a positive impact on participants' confidence in science.

The difference between participants and comparator pupils was not statistically significant. Participants saw a smaller increase than the national average (+3.5 percentage points).

Table 7

	Type of Pupils	Sample size	Baseline	Endline	Difference	Percentage point difference	Statistical significance
Self-efficacy	Comparators	26	4.91	4.90	-0.01	-0.07%	$p = 0.67$
	Participants	29	4.75	4.89	+0.14	2.30%	

	Benchmark	Refer to methodology	4.66	4.87	+0.21	3.50%	

This was corroborated by tutors and teachers at both Ernest Bevin and Burntwood. A tutor at Ernest Bevin stated "*they kind of gained confidence as the year went on*". Similarly, a tutor at Burntwood said:

> *"the biggest difference I really noticed in them was definitely their confidence"*.

Teachers from both schools agreed with the assessment that pupils' confidence in science had increase. The teacher from Ernest Bevin stated:

> *"I saw kids in Science Stars putting their hands up a lot more towards the end, you know, asking me questions around the subject or interweaving questions about different topics."*

The teacher from Burntwood re*ported:*

> *"some of them switched from Foundation to Higher even, because it boosted their confidence."*

The teacher from Burntwood provided a compelling example of how the Science Stars programme impacted a specific pupil's confidence:

> *"So, for example, I had one of our students, she's in Step 5, she is bright, but because of the language barrier, she's EAL, she, you know, she clearly needed that time, that intervention, where she had time to understand the keywords, break down exam questions, for example, which was given by the tutors. They had this time where they put in exam questions, they're going through them, which is not always easy to do every lesson for us as teachers. So, you know, she in the lessons that she started, you know, putting up her hands more, she became really confident, she was, before it'd be her like sort of one word answers, you know, she could put things into sentences, she was saying keywords, and then she was the one that went from foundation to hire. She really built that confidence in herself. So I think there was, you know, there were, I think for her, she really made the most out of the programme*

Key finding: On average, participants decreased their levels of test anxiety (-3.92 percentage points) more than their comparator peers (-1.43 percentage points), suggesting that Science Stars had a positive impact on participants' anxiety around tests and exams.

The difference between participants and comparator pupils was not statistically significant.

Table 8

	Type of Pupils	Sample size	Baseline	Endline	Difference	Percentage point difference	Statistical significance
Test anxiety	Comparators	35	3.69	3.60	-0.09	-1.43%	$p = 0.7$

	Participants	34	3.46	3.23	-0.24	-3.92%	

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, but the differences are not statistically significant.

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

Table 9

Non-cognitive skill	Type of Pupils	Sample size	Baseline	Endline	Difference	Percentage point difference	Statistical significance
Metacognition	Comparators	15	4.09	4.29	+0.20	3.33%	$p = 0.22$
	Participants	8	3.95	4.80	+0.84	14.02%	
Self-efficacy	Comparators	12	4.55	4.89	+0.34	5.71%	$p = 0.52$
	Participants	12	4.58	5.34	+0.76	12.65%	
Test anxiety	Comparators	17	3.84	3.78	-0.06	-0.98%	$p = 0.22$
	Participants	14	4.12	3.31	-0.81	-13.49%	

This data suggests the Science Stars programme is having a more positive impact on non-cognitive and SEMH skills in Ernest Bevin than those at Burntwood. Tutors who delivered the Science Stars programme at Ernest Bevin corroborated this when reporting they had taught revision specific revision skills and had seen a positive difference in pupils.

> *"I taught them how to do keyword type notes which was more time or so helps them more in space repetition"*
>
> *"They were using mnemonics in their actual revision as well"*
>
> *"I saw a difference when I started encouraging them to revise beforehand"*

The teacher at Ernest Bevin also reported on improved revision skills in their pupils:

> *"More of the Science Star students were asking me for past paper questions than any other of my students".*

The same teacher also remarked that Science Stars was having a profound impact on revision skills and awareness of their own strengths and weaknesses, saying that it was…

> *"instilling something that they were very weak at in terms of independent work and then making that independent work something that was more routine to them"*

The teacher at Ernest Bevin mentioned that the programme had even had positive impacts on pupils' confidence who were not participating in the programme, reflected in the positive increase in the comparator group's self-efficacy score:

> *"It didn't only just increase the kids' confidence that were in Science Stars, but it made the other kids around them have increased confidence too."*

When it came to tutors and teachers at Ernest Bevin reporting on anxiety around tests and exams, there was one tutor who indicated that some of their pupils saw a reduction in anxiety – *"I think at the start they showed quite a lot of anxiety… they got their mock results, two of them were very happy"*.

However, it was made clear that was not representative of all pupils' anxiety levels: *"it was the students who were working at a lower level that were anxious, the one kid who was working at a bit of a higher level never expressed any worries or concerns"*

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 change rather than a true effect.

The broad picture for Burntwood pupils' non-cognitive and SEMH skills deviates from the overall trend initially described. Participants slightly improved their metacognition (+1.3 percentage points), but there was less improvement than the comparator group (+2.7 percentage points). This change was not statistically significant. Their self-efficacy decreased slightly less (-5.03 percentage points) than their peers (-5.01 percentage points); this difference was not statistically significant. Participants' test anxiety went up slightly (+2.78 percentage points) whereas the comparator groups' test anxiety decreased (-1.85 percentage points). The difference between these two groups was not statistically significant.

Table 10

Non-cognitive skill	Type of Pupils	Sample size	Baseline	Endline	Difference	Percentage point difference	Statistical significance
Metacognition	Comparators	17	4.69	4.81	+ 0.12	2.07%	$p = 0.88$
	Participants	18	4.34	4.41	+ 0.07	1.30%	
Self-efficacy	Comparators	14	5.21	4.91	- 0.3	-5.03%	$p = 1$
	Participants	17	4.88	4.58	- 0.3	-5.01%	
Test anxiety	Comparators	18	3.54	3.43	- 0.11	-1.85%	$p = 0.57$
	Participants	20	3.00	3.17	+0.17	2.78%	

Burntwood tutors reported improvements in students' revision skills and self-learning awareness, though feedback varied across the group of tutors.

Tutors referenced that *"they were using the keywords"*. One tutor, however, stated that they *"didn't see any evidence of any change in revision techniques"* reflecting the trend that participants' metacognition fared worse than their comparator peers. The teacher at Burntwood stated that there was some improvement in pupils' revision skills because "*they've got their revision books, they've got their flashcards*".

Tutors and teachers both corroborated an increase in anxiety in the pupils participating in the Science Stars programme:

"I think there was more nerves towards the end of the programme because the exams were nearby." - *Tutor*

"everyone's got a bit of anxiety before an exam, don't they? So in that sense, you know, they are all quite nervous." - *Teacher*

Attainment – All pupils

Key finding: On average, participants increased their science grade from 4.14 to 4.69 (+0.54) which was more than the progress made by comparators (3.83 to 4.12, +0.29). This difference was not statistically significant (p = 0.13, n = 68).

The absence of statistical significance may be attributed to insufficient statistical power resulting from the small sample size. The differences above may also be better explained by random chance rather than there being a genuine difference between the two populations.

In either case, participants demonstrated greater improvement in attainment between the pre- and post-programme assessments compared to their peers in the comparator group. This indicates potential benefits the programme may have provided the participants in their understanding of Science GCSE content. This improvement is particularly noteworthy given that participants started with higher baseline grades than their peers in the comparison group. Such progress is compelling, as students with higher initial scores typically have less room for improvement compared to those starting at lower levels.

Table 11

Attainment	Type of Pupils	Sample size	Baseline – Autumn Mock	Endline - GCSE	Difference	Percentage point difference	Statistical significance
Science	Comparators	33	3.83	4.12	0.29	3.60%	p = 0.13
	Participants	35	4.14	4.69	0.54	6.79%	

Key finding: 83% of participants achieved their target grade which was more than the 67% of comparator peers who reached their target. This difference between participants and comparators was not statistically significant (p = 0.13, n = 68).

The absence of statistical significance in the two findings above may be attributed to insufficient statistical power resulting from the small sample size. The differences above may also be better explained by random chance rather than there being a genuine difference between the two populations.

This data points to the likely contribution the Science Stars programme is having on participants' understanding of Science, their ability to answer exam questions and their academic attainment.

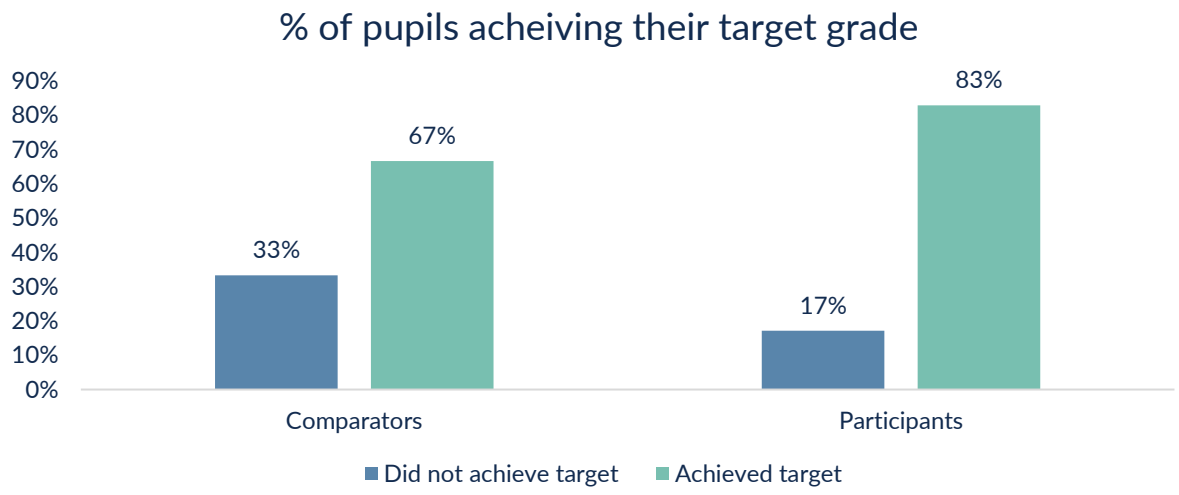


Figure 1

Attainment – Ernest Bevin

Key finding: The progress made in GCSE science was more positively pronounced on participants in Ernest Bevin than in the overall picture. On average, their grades increased from 3.53 to 4.47 (+0.95), whereas their comparator peers' grade increased less (3.18 to 3.38 (+0.21 grades); this difference was statistically significant ($p < 0.05$, n = 36).

This trend, of positive impact being more pronounced in Ernest Bevin pupils, is like the one seen in the non-cognitive and SEMH skills. Ernest Bevin's longer engagement with the program has likely led to more deeply embedded processes, contributing to these more pronounced observable changes.

Table 12

Attainment	Type of Pupils	Sample size	Baseline – Autumn Mock	Endline - GCSE	Difference	Percentage point difference	Statistical significance
Science	Comparators	17	3.18	3.38	0.21	2.57%	$p < 0.05$
	Participants	19	3.53	4.47	0.95	11.84%	

This level of increase in grade reflects the narrative provided by tutors and teachers at Ernest Bevin that pupils were better at understanding how to answer exam questions and had greater understanding of Science GCSE content.

> *One tutor noted: "they were able to answer questions from the previous week which was quite rewarding actually thinking that they actually did go away and do that."*
>
> *Another noted: "They did get more confident with the content".*
>
> *A teacher stated that "kids started understanding the command words a lot better, and how the mark system worked".*

Both tutors and teachers at Ernest Bevin provided specific examples of pupils' increase in Science GCSE knowledge

"one kid was able to identify that electrons have a negative charge and protons have a positive charge which is [...] very Foundation things but it was an improvement from when I first met them."

A teacher reflected that:

"when questions pop up about animal plant cells, they had the knowledge to answer it. Even when it got to harder levels where it was like mitosis, you know, and kids were telling me like, oh yeah, blood cells, red blood cells don't have a nucleus itself, so they don't do, you know, it was, it was like the base knowledge carried them in certain topics."

Key finding: 100% of participants at Ernest Bevin achieved their target grade (in comparison to 94% of their comparator peers who achieved this grade.) This difference was not statistically significant (*p* = 0.3, n = 36)

This small difference in the percentage of participants and comparator peers achieving their target grade shows that all pupils at Ernest Bevin are generally well-supported in achieving their aims. Given Ernest Bevin's more conservative approach to setting target grades compared to Burntwood, the higher proportion of students achieving their targets should be interpreted within this context. The difference in target-setting practices between the two schools affects how their respective achievement rates should be interpreted.

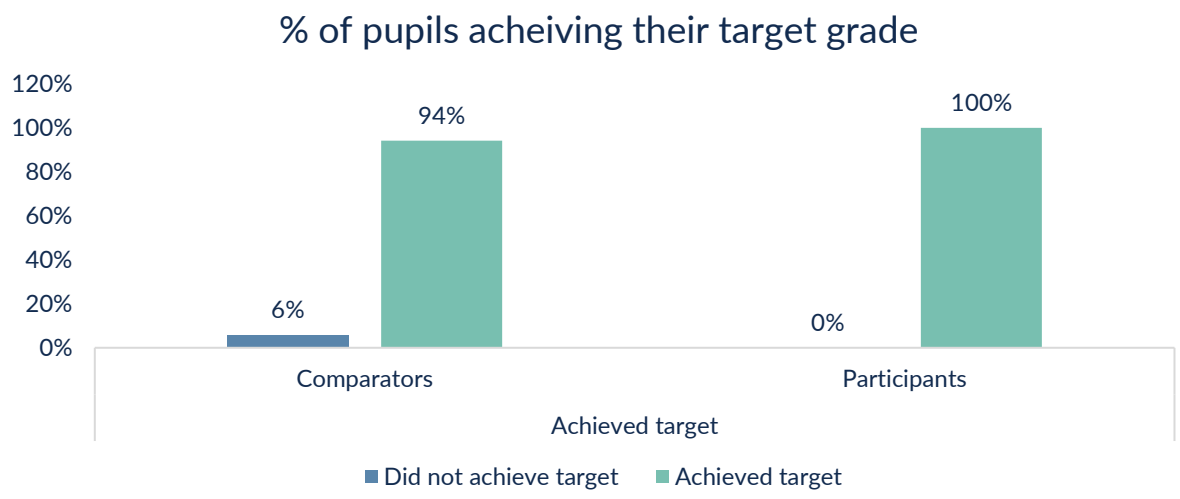


Figure 2

Attainment – Burntwood

Key finding: Participants at Burntwood saw an increase in their grade from Autumn mock to their GCSE (+0.06 grades) but this was less of an improvement than their comparator peers (+0.38 grades). This suggests that participants' ability to answer exam questions or their understanding of Science GCSE did not improve. This difference was not statistically significant.

Although this shows that comparator pupils are faring better when it comes to improving academic achievement and understanding their knowledge of Science, it should be noted that this result is not statistically significant (p = 0.39, n = 32). The absence of statistical significance may mean the difference may be better explained by random chance rather than there being a genuine difference between the two populations.

Despite their smaller rate of improvement compared to their peers, Burntwood participants' final grades surpassed their Burntwood comparator peers.

Tutors noted they had seen an increase in their pupils' ability to answer exam questions, with one tutor highlighting specific approaches that worked in developing pupils' exam answering techniques:

"After doing those diagrams and drawings on the board, they understood it and then they were able to answer the questions more confidently".

One tutor also identified a specific area where they saw their pupils' content knowledge increase:

"They didn't know the menstrual cycle, they found a bit confusing. So, I found [that] drawing out together with them and going [through] the steps together and visually putting [it] on the board, even though the diagram at the end didn't look very good, it helped them a lot and they said that it was really helpful to [draw it out] with them."

Table 13

Attainment	Type of Pupils	Sample size		Baseline – Autumn Mock	Endline - GCSE	Difference	Percentage point difference	Statistical significance
Science	Comparators	16		4.53	4.91	0.38	4.69%	p = 0.39
	Participants	16		4.88	4.94	0.06	0.78%	

Key finding: A greater percentage of participants (63%) achieved their target grade in comparison to their comparator peers (38%). This was not statistically significant (p = 0.17, n = 32).

This finding indicates Science Stars is having a positive impact on participants' ability to reach the standard expected from them. This insight helps provide context for understanding why participants' academic progress appears to be developing at a different rate than their peer group.

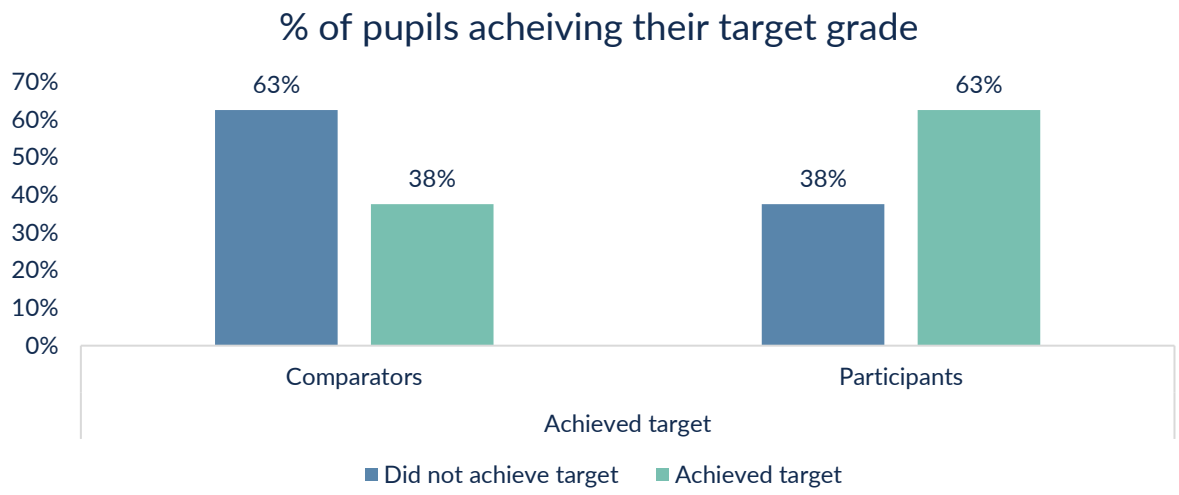


Figure 3

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 all these three measures this year, however, was smaller than the progress made last year.

Although participants' positive progress is clear, it is worth investigating why the rate of progress has slowed down from last year. One school was new to the programme this year, which may explain the lower impact. Unlike the established school, this new school had less time to fully implement and embed the programme's practices.

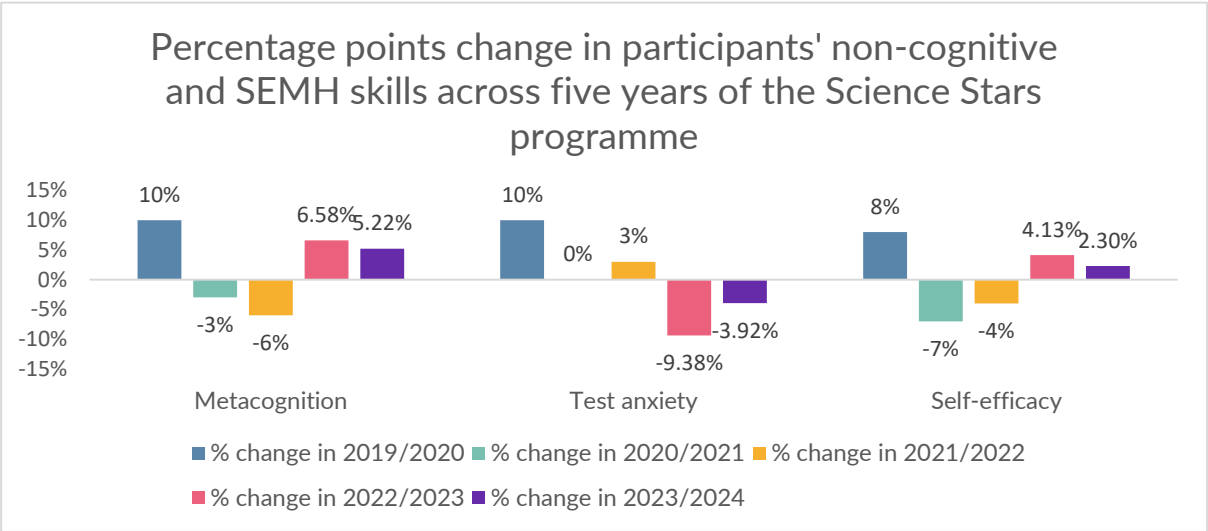


Figure 4

Cross-year: Attainment – all pupils

Key finding: 2023/24 saw a larger percentage point increase in Science grades than in 2022/23 and reached similar levels to previous years of the Science Stars programme.

Considering that from 2021/22 to 2022/23 saw a decrease from a 15-percentage point increase to just a 5.38 percentage point increase in participants' Science grade, it is positive to see the percentage point increase to return to slightly above 15.

Table 14

Attainment	Type of Pupil	% change in 2019/2020	% change in 2020/2021	% change in 2021/2022	% change in 2022/2023	% change in 2023/2024
Science	Comparators	2%	2%	16%	-0.38%	13.88%
	Participants	10%	15%	15%	5.38%	15.02%

Key finding: The difference between participants' and comparator peers' percentage point progress made from their Autumn term mock to their final GCSE grade decreased from 2022/23 to 2023/24.

In the academic years of 2019/2020 and 2020/21, there was a marked positive difference between the rate of progress for participating pupils versus comparator pupils. This then was reversed in 2021/22, where comparator peers achieved greater progress than their participating peers. This was flipped again in 2022/23, where participants once again saw greater progress than comparator peers. Although this trend has continued this year, it is less marked than last year.

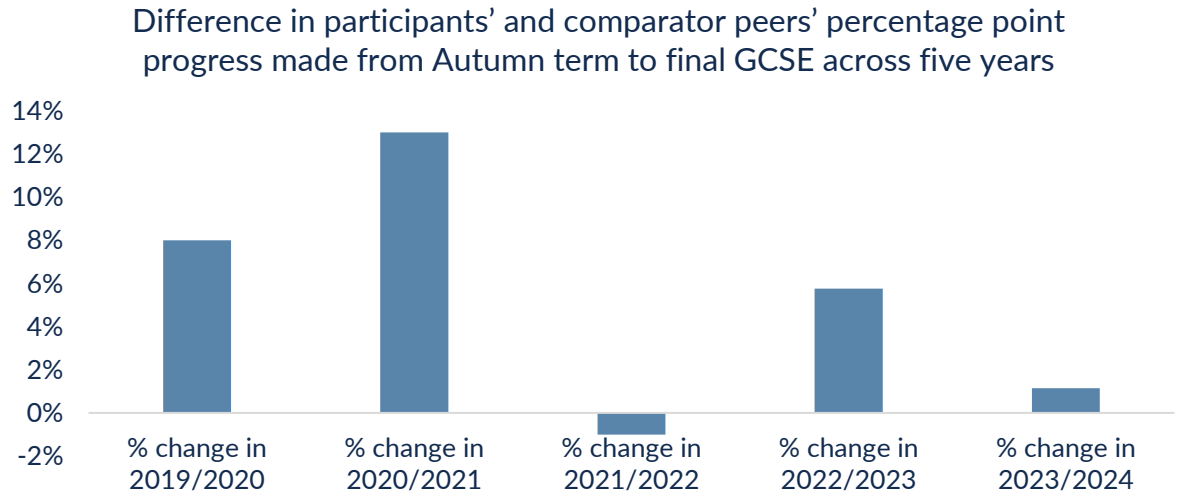


Figure 5– A positive percentage on this graph indicates participants improving their Science grade more than their comparator peers. A negative percentage change shows the opposite (comparator pupils improving more than participants).

Key finding: The percentage of participants achieving their target grade has increased by over 20 percentage points from last year to this year.

This data suggest that Science Stars programme could be having a positive effect on pupils' achieving their target grade.

Table 15

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

Key finding: The positive difference in the percentage of pupils achieving their target grade between participating pupils and their comparator peers has slightly decreased from 2022/23 to 2023/24.

It should be noted that this decrease was only slight and the picture from 2020/21 to 2023/24 shows that around 20% more of pupils in the participating group achieved their target grade in comparison to their comparator peers. This difference was much less pronounced in 2021/22.

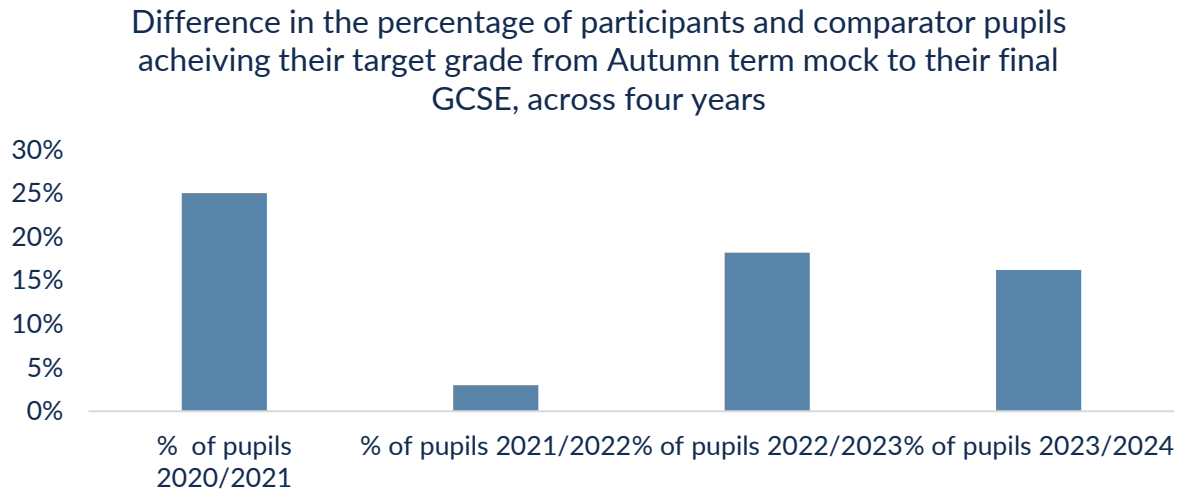


Figure 6 – A positive percentage on this graph indicates a greater proportion of participants achieving their target grade than their comparator peers.

Programme Delivery

The Science Stars' programme received feedback from both tutors and teachers, highlighting its successes while identifying areas for improvement.

Tutors' Motivation

Tutors at Ernest Bevin were motivated to become tutors because of their previous positive encounter with tutoring, wanting to teach in smaller group setting, and finding satisfaction in being role models. They were also motivated by transforming their personal educational journeys into opportunities for pupils. They also noted that they wanted to promote effective learning practices and wanted inspire confidence and academic progress in Science Stars students.

Tutors at Burntwood were motivated to become tutors because of the emotional reward from seeing pupils' progress. Some tutors noted that their passion for teaching had motivated them to participate in the programme. They appreciated that the role was flexible and would provide them with teaching experience, which may be helpful for employment in the future.

Successes and Areas of Improvement of the Science Stars programme

Tutors

Tutors from both Ernest Bevin and Burntwood highlighted that smaller group setups improved one-on-one interactions and engagement. Tutors reflected that the initial training was beneficial and that specific teaching tactics, such as drawing complex processes, facilitated better understanding among students.

Training content was sometimes too basic for returning tutors but was also noted as overly detailed by other tutors. Communication issues included inconsistent updates about student absences and challenges aligning with the school calendar. Managing mixed-ability groups and varying student motivation was difficult, with calls for more training in these areas

Tutors recommended the following as elements to improve for the following year's programme:

- Tailor teaching support, such as grouping students by academic ability and integrating diverse learning skills into teaching methods.
- Enhanced communication processes, including better coordination with school staff and alignment with school schedules.
- More structured training to better address the specific challenges of tutoring.

Teachers

A teacher noted that effective student selection for the programme contributed to the programme's success. Another teacher noted that the school's partnership with St George's has strengthened pupil relationships.

Teachers also noted that there had been some administrative issues, including difficulties with handovers from the previous school teacher leading on the programme. They also noted

issues with questionnaire distribution, specifically around knowing how to indicate whether a pupil was a participant or part of the comparator group. They also noted some concerns for pupil safety with the programme ending in the evening. One of the teachers noted issues with the quick start of the programme and hoped there would be a smoother start in the academic year 2024/25.

Teachers recommended the following:

- Smoother programme transitions to avoid rushed starts.
- Addressing logistical and safety concerns, particularly for students traveling at night.

4. Conclusion and Recommendations

In 2023/24, the Science Stars programme continued to demonstrate impact across participant schools. Overall, participants showed improvements in non-cognitive and SEMH skills, with generally positive trends in Science attainment compared to their comparator groups.

At Ernest Bevin, the results were consistently positive. However, Burntwood School presented a more nuanced picture. While Burntwood participants saw some improvements, their outcomes were more mixed. Specifically, they experienced challenges in several key areas: participants showed less improvement in metacognition compared to their comparator peers, experienced a decrease in self-efficacy, and saw an increase in text anxiety, whereas the comparator pupils saw a decrease. In terms of Science attainment, Burntwood participants made smaller progress relative to their comparator peers. Nevertheless, more Burntwood participants achieved their target grades compared to the comparator group.

The 2023/24 data continues the broader trend of programme participants progressing in their non-cognitive and SEMH skills and progressing more than their comparator peers. However, the positive progress was less pronounced compared to the previous year's (2022/23) results.

Based on the findings, the following recommendations are presented:

Evaluation Recommendations

- Investigate the differential impact on the two schools' pupils' non-cognitive and SEMH skills.
- Investigate differential impact on academic progress.
- Analyse external factors that may have affected pupils' lower gains in SEMH and non-cognitive skills compared to the previous year.
- In 2024/25, Year 11 pupils' target grades will be constructed differently to previous years due to covid-19 having impacted their Year 6 SATs in 2019/20; St George's and ImpactEd Evaluation to discuss how this may impact future evaluations.
- Consider a more rigorous matching procedures so more robust conclusions can be drawn between participating and comparator pupils.

Programme and Delivery Recommendations

- Consider incorporating additional stress and anxiety management workshops.
- Teachers and tutors could share best practices among their peer groups to maximize programme impact.
- Tailor teaching support, such as grouping students by academic ability and integrating diverse learning skills into teaching methods.
- Enhanced communication processes, including better coordination with school staff and alignment with school schedules.
- Ensure that all training is relevant and implementable by student teachers.
- Ensure that all trainings are attended by all students teachers.

- Ensure a system for early planning of programme to ensure schools feel confident establishing the programme.
- Acknowledge and address the logistical and safety concerns, particularly for students traveling at night.

Appendix

Glossary

Only terms relevant to the evaluation report should be included in the 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.

Comparator group

A comparator 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.

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 programme's 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.

Education terminology

EAL

Pupils with English as an Additional Language (EAL) refers to learners whose first language is not English.

Pupil Premium (PP)

The pupil premium grant is designed to allow schools to help disadvantaged pupils by improving their progress and the exam results they achieve. Whether a child is eligible for Pupil Premium funding is often used by schools as an indicator of disadvantage.

Measures for social and emotional skills

The self-report measures available on the ImpactEd platform are academically validated questionnaires for measuring 'social and emotional' skills that have the biggest impact on pupil life chances and outcomes.

Metacognition

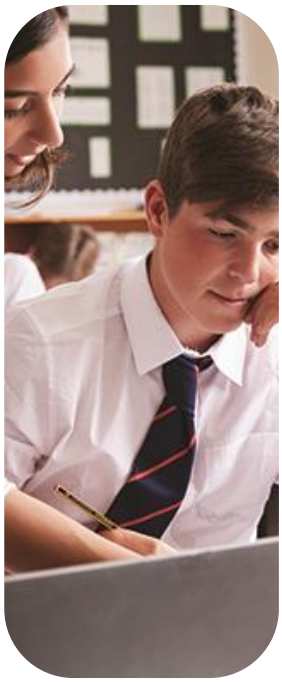
Metacognition means 'thinking about thinking': pupils' ability to think explicitly about their own learning (Flavell, 1979; Higgins et al., 2016). It is strongly associated with academic progress and improves other skills required for learning, such as critical thinking. Metacognition enables pupils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

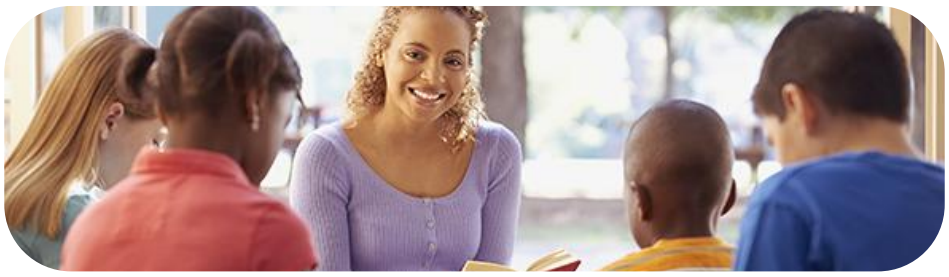
Self-efficacy

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.





Supporting our purpose driven partners to make better decisions using high quality evidence.

Get in touch

hello@impacted.org.uk

www.evaluation.impactedgroup.uk

ImpactEd Evaluation

ImpactEd Group