Science Self-Testing Toolkit

Pilot Report

17th July 2020

Arnaud Vaganay, Sarah Frankenburg, Molly Mayer, Helen Burridge, Bethany Thompson, and Anysia Nguyen.

The Education Endowment Foundation (EEF) is an independent grant-making charity dedicated to breaking the link between family income and educational achievement, ensuring that children from all backgrounds can fulfil their potential and make the most of their talents.

The EEF aims to raise the attainment of children facing disadvantage by:

- identifying promising educational innovations that address the needs of disadvantaged children in primary and secondary schools in England;
- evaluating these innovations to extend and secure the evidence on what works and can be made to work at scale; and
- encouraging schools, government, charities, and others to apply evidence and adopt innovations found to be effective.

The EEF was established in 2011 by the Sutton Trust as lead charity in partnership with Impetus Trust (now part of Impetus - Private Equity Foundation) and received a founding £125m grant from the Department for Education. Together, the EEF and Sutton Trust are the government-designated What Works Centre for improving education outcomes for school-aged children.

The project is jointly funded by the Wellcome Trust as part of our Improving Science Education Round.

For more information about the EEF or this report please contact:

Jonathan Kay
Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
SW1P 4QP

0207 802 1653

jonathan.kay@eefoundation.org.uk

www.educationendowmentfoundation.org.uk

About the evaluator

The project was independently evaluated by a team from NatCen Social Research: Arnaud Vaganay, Sarah Frankenburg, Molly Mayer, Helen Burridge, Bethany Thompson, and Anysia Nguyen.

The lead evaluator was Arnaud Vaganay.

Contact details:

Arnaud Vaganay
NatCen Social Research
35 Northampton Square
London EC1V 0AX

Email: **Arnaud.Vaganay@natcen.ac.uk**

Tel: 0207 549 9570

Table of contents

List of tables

Executive summary

The project

The core intervention evaluated in this report is a Science Self-Testing Toolkit (SSTT). The developers of SSTT were three Research Schools—Kingsbridge, Durrington, and Huntington.[1] These schools acted as local 'hubs'. Each hub recruited three other schools to deliver the SSTT bringing the total number of pilot schools to 12. 2,100 students took part in the pilot. The intervention was delivered in school by science teachers, led by heads of science, and supported at home by parents.

SSTT is a suite of five evidence-informed, content-free strategies to be used and deployed by teachers, students, and parents at Key Stage 4 (KS4). These five strategies are:

- pre/post tests;
- flashcards;
- mind-mapping tests;
- structured note-taking; and
- cumulative quizzing.

These approaches encourage students to engage in active retrieval of knowledge throughout the curriculum, as well as during revision. The implementation of the intervention was facilitated by a number of support activities including two external training days held by hub schools, in-school short training sessions (‘cascading’), in-school coaching, after-school briefings and updates for teachers (‘twilight sessions’), and parent engagement sessions.

The aim of this study was to evaluate the implementation of the intervention. Our research explored the profile of participants, participants’ usual practice, evidence of promise, the feasibility of integrating the SSTT into the Year 10 science curriculum, the feasibility of delivering support activities, the cost of the project, and the intervention’s readiness for trial. The original evaluation protocol was significantly modified to make the evaluation proportionate to the scale of the intervention and reduce the burden on participants. We collected data by means of interviews, observations, desk research, and an online student survey. The pilot ran between January and July 2019. The project is jointly funded by the Wellcome Trust as part of our Improving Science Education Round.

Key conclusions

Research question	Finding
Is there evidence to support the theory of change?	Students responded well to the intervention, but some strategies were perceived as more effective and enjoyable than others. Students with higher science self-efficacy had a more positive experience than students with lower science self-efficacy. Some strategies in the SSTT are close to usual practice (flash card revision).
Is the approach feasible to deliver?	Heads of science and science teachers responded well to the intervention, both in terms of attitudes and behaviours. Support activities were faithfully implemented and the cost of SSTT is low.
Is the intervention ready to be evaluated in a trial?	Despite the above-mentioned strengths, further consideration must be given to a few issues to make a trial feasible as the conditions of implementation were more favourable than they would be in a trial. These issues include (1) the innovativeness of the intervention, (2) clarity over the target year group and optimal duration of the programme, (3) the frequency and use of the SSTT, (4) the developers’ willingness and capacity to recruit and deliver on a greater scale, and (5) the responsiveness of schools and teachers to the evaluation.

[1] The Research Schools Network is a collaboration between the EEF and the Institute for Effective Education (IEE) to fund a network of schools that supports the use of evidence to improve teaching practice. Research Schools work with the other schools in their areas to help them to make better use of evidence to inform their teaching. For more information, see https://educationendowmentfoundation.org.uk/scaling-up-evidence/research-schools/

Additional findings

Overall, almost two thirds (62%) of students said they want to keep using the SSTT in class. A similar number of respondents (64%) said it will be helpful to understand and revise other subjects. Flash cards and structured note-taking were perceived as the most effective (in terms of memorisation and understanding) and enjoyable strategies. Cumulative quizzing was viewed as the least effective and enjoyable strategy.

Students with higher 'science self-efficacy' had a more positive experience of the SSTT than students with lower science self-efficacy. This could mean that there is a risk of attainment gaps widening. Male and female students had similar experiences of the SSTT. Students and teachers indicated that they were familiar with some of the strategies in the SSTT even if they had not used them in a 'self-testing' way (for example, flash cards).

Overall, heads of science and teachers responded well to the intervention. Teachers understood the benefits of self-testing and found the toolkit easy to use. Although we noted some reservations about aspects of the intervention (including its effect on low prior-attaining students), none of the interviewees expressed major concerns. There were also positive reported changes to teachers' behaviour. They reported using the SSTT on a regular basis, sometimes in very creative ways. Some teachers successfully introduced the SSTT in Year 9 thereby exceeding the developers' expectations. However, it should be noted that (1) students reported using the SSTT less frequently than the teachers did and (2) the reported frequency of use varied significantly between students. This variation suggests that the expected dosage was unknown or not considered optimal.

With a few exceptions, the developers delivered support activities as planned. When a support activity did not take place, support was generally provided in a different way or at a different time. Overall, teachers felt supported by the developers, enjoyed the training and coaching, and found support activities a good use of their time. Some teachers cascaded their learning to the school's leadership and to teachers in other departments thereby exceeding the developers' expectations.

Despite a sound logic model, a high level of implementation fidelity from the developers, a positive response among both teachers and students, and a low cost, our conclusion is that further consideration must be given to a few key issues before a trial can be deemed feasible. These issues include (1) the innovativeness of the intervention, (2) clarity over the target year groupand the optimal duration of the trial, (3) the optimal frequency and use of the intervention, (4) the developers' willingness and capacity to recruit and deliver on a greater scale, and (5) the responsiveness of schools and teachers to the evaluation.

Many teachers and students told us that they had already used some of the SSTT tools before the pilot. While teachers recognised the benefits of the 'self-testing' approach and of combining these tools into a single kit, it was not clear how innovative they found the SSTT. The insufficient contrast between intervention and control reduces the likelihood of detecting an impact.

NatCen's assessment is that the conditions of implementation were more favourable than they would be in a trial, for example, because of the small total number of schools involved in the pilot. In addition, the individuals interviewed as part of the evaluation are likely to be more supportive of the SSTT than teachers who were not interviewed.

1. Introduction

1.1. Intervention

GCSE science exams have recently changed from modular exams throughout Years 10 and 11 to final exams at the end of Year 11. This has significant implications for students who are now required to memorise large amounts of content over the two-year course.

The intervention evaluated in this report sought to address this challenge through self-testing. This is a teaching and learning technique that encourages students to engage in active retrieval of memories rather than more passive approaches such as re-reading material.

The core intervention evaluated in this report is a Science Self-Testing Toolkit (SSTT). SSTT is a suite of five evidence-informed, content-free strategies to be used by teachers and students at Key Stage 4 (KS4). They are:

- **Pre-tests/Post-tests.** Doing a test before you have learned the lesson or topic. Once you have learned the lesson or topic, you do the test again.
- **Cumulative quizzing**. Coming up with questions about a topic you are learning and adding to the list of questions over time, testing yourself on each question each time. Your teacher may give you the questions, you may make them up as a class, or you make them up on your own.
- **Flashcards.** Making flashcards (either on paper, the computer, or an app). Over time, you shuffle the order of the deck of cards and remove some cards for a few days before adding them back in. Your teacher may give you the flashcards, you may make them as a class, or you may make them up on your own.
- **Structured note-taking.** Writing down questions as you write or read notes in the margin of your page and then trying to answer the questions without looking at your notes. Your teacher may give you the questions, you may make them up as a class, or you make them up on your own.
- **Blind mind mapping.** Creating a mind map for a topic you have learned, tracing the structure of the map on a blank piece of paper, and attempting to fill in the information whilst looking at your notes as little as possible. Your teacher may give you structured mind map, you may make it up as a class, or you make it up on your own.

The developers assumed that many teachers would be familiar with some of these strategies, such as flash cards and mind-mapping. They felt that the novelty was in the 'self-testing' element where students produce both the question and the answer rather than the usual practice where questions and answers are produced by different people. Self-testing is one of the learning techniques reviewed by Dunlosky et al. (2013) and the EEF in the **Metacognition and Self-Regulated Learning Report** (Quigley, Muijs, Stringer, 2018). In addition, the developers felt that bringing together these strategies in a single toolkit would encourage their take-up and creative use.

The implementation of the intervention was facilitated by a number of **support activities**, including:

- two days of training provided by the developers to the head of science and one science teacher from each of the pilot schools;
- cascading activities—condensed training sessions delivered by the trainees themselves to their colleagues;
- in-school coaching provided by the developers to heads of science and science teachers;
- twilight sessions—briefings and updates organised in pilot schools by the developers for heads of science and teachers; and
- parent engagement sessions.

The developers of SSTT were the Kingsbridge, Durrington, and Huntington Research Schools. The three Research Schools acted as local 'hubs'. Each hub recruited three other schools to deliver the SSTT bringing the total number of pilot schools to 12. The intervention was delivered in school by science teachers and led by heads of science.

The intervention was expected to benefit all children in participating schools' Year 10 science classes.

The EEF and the Wellcome Trust co-funded both the delivery of the intervention and its evaluation.

The pilot ran in 12 schools between January and July 2019.

1.2. Background evidence

The effect of self-testing on learning outcomes is well documented. Overall, this effect seems to be positive. A 2017 meta-analysis summarising 118 articles (272 effect sizes, N = 15,000) found that it is, on average, more beneficial than any other learning strategy when the practice test and the final test are based on the same format or take place in identical conditions, for example, in class (g = 0.61, 95% CI: 0.58, 0.65; Adesope, Trevisan, Sundararajan, 2017). A 2018 meta-analysis summarising 67 articles (192 effect sizes, N = 10,000) found that it is, on average, more beneficial than any other learning strategy even when the practice test and the final test are based on different formats or take place in different conditions, for example, first at home and then in class (d = 0.40, 95% CI: 0.31, 0.50; Pan, Rickard, 2018).

However, the effect of self-testing on attainment depends on a range of factors. According to a 2017 meta-analysis, the effect of self-testing on learning outcomes seems to be stronger (a) when the initial learning involves reading or studying a passage, rather than listening, (b) for mixed-format practice tests (that is, including a mix of free-recall, cued-recall, and short-answer tests) than for practice tests using a single type of test, (c) for secondary school students than for students at other levels, (d) when the practice and final tests formats are identical, although this point is contested (see below), (e) when the time lag between practice and final test is between one and six days, and (f) for high treatment fidelity studies. On the other hand, the effect of self-testing appears broadly similar (a) with or without feedback and (b) when the final test is administered in the class room or in a lab (Adesope, Trevisan, Sundararajan, 2017).

More research is needed on the possible adverse effects of self-testing. First, some experts have warned that an over-utilisation of tests (including self-tests) in school curricula can result in superficial (or shallow) learning. However, there is no evidence that self-testing encourages shallow learning. Second, testing can also have a negative impact on students' emotional health and wellbeing. The World Health Organisation (Currie et al., 2012) found that 11- and 16-year-old students in England feel more pressured by their school work than in most other European countries. McCaleb-Kahan and Wenner (2009), drawing on research in the U.S.A., report that, as the number and the importance of tests used in schools has increased, the number of students who experience test anxiety has also increased. However, here again, there is no direct evidence of a link between self-testing and anxiety.

1.3. Research questions

The objective of this study was to assess the following implementation and process evaluation (IPE) dimensions:

- Evidence of promise—the extent to which the intervention delivered its main outcomes at the pilot stage. The primary outcome of the pilot was the self-reported effect of the intervention on memorisation and understanding.
- Feasibility—the sum of all drivers and obstacles to the success of the intervention at the pilot stage.
- Scalability—the likelihood for the intervention to produce the same outcomes as in the pilot when scaled up.

1.4. Ethical review

NatCen's Research Ethics Committee (REC) reviewed and approved the research proposal for this project on 11 January 2019. The committee consists primarily of senior NatCen staff. The final study plan incorporates the guidance and recommendations provided by the REC.

Legal basis for processing personal data

NatCen was the data controller and processor for this evaluation. The legal basis for processing personal data is covered by GDPR Article 6 (1) (f): '**Legitimate interests: the processing is necessary for your (or a third party's) legitimate interests unless there is a good reason to protect the individual's personal data which overrides those legitimate interests.**'

Our assessment was that the evaluation fulfilled one of NatCen's core business purposes (undertaking research, evaluation, and information activities) and was therefore in our legitimate interest, and that processing personal information was necessary for addressing the research questions in this study. We considered and balanced any potential impact on the data subjects' rights and found that our activities would not do the data subject any unwarranted harm.

1.5. Delivery team

The delivery team included:

- at Kingsbridge Research School: Lorwyn Randall (strategic lead—overseeing programme design, resource development, workshop delivery, and follow-on support) and Jon Eaton (project lead—workshop facilitation and coaching support);
- at Huntington Research School: Jane Elsworth (strategic lead—programme design team, resource development, workshop delivery and follow-on support), and Penny Holland (project lead—workshop facilitation and coaching support); and
- at Durrington Research School: Shaun Allison (strategic lead—programme design team, resource development, workshop delivery, and follow-on support), and Steph Temple (project lead—workshop facilitation and coaching support).

1.6. Evaluation team

The evaluation team included:[2]

Conceptualisation	Arnaud Vaganay
Data curation	Sarah Frankenburg, Molly Mayer, Helen Burridge, Bethany Thompson
Analysis	Arnaud Vaganay, Sarah Frankenburg, Molly Mayer, Helen Burridge, Bethany Thompson, Anysia Nguyen
Funding acquisition	Arnaud Vaganay
Investigation	Sarah Frankenburg, Molly Mayer, Helen Burridge, Bethany Thompson
Methodology	Arnaud Vaganay
Project administration	Arnaud Vaganay, Sarah Frankenburg
Resources	NatCen Social Research
Software	Sarah Frankenburg, Anysia Nguyen, Molly Mayer
Supervision	Arnaud Vaganay
Validation	Arnaud Vaganay, Ellen Broome
Visualisation	N/A
Writing—original draft	Arnaud Vaganay, Helen Burridge, Bethany Thompson, Anysia Nguyen
Writing—review and editing	Arnaud Vaganay, Helen Burridge

All evaluators were affiliated with NatCen Social Research at the time of writing.

[2] Based on the CRediT taxonomy of research roles: https://casrai.org/credit/

2. Methods

2.1. Study plan

The study plan for this evaluation was published in February 2019.[3] We substantially modified this during the course of the project to reduce the burden on schools and to reflect the occasional unresponsiveness of some informants and gatekeepers. The IPE dimensions presented in section 1.3 were also modified at the reporting stage to maximise readability (see section 2.7). The section below provides a brief overview of the evaluation design, both as initially planned and as executed. All post-design changes were discussed and agreed with the EEF.

2.2. Recruitment

SSTT was piloted in three areas of England: South Devon, Yorkshire, and West Sussex. In each area, a Research School acted as local hub ('hub school') and recruited three local schools to run the pilot alongside them—a total of 12 pilot schools. Schools were recruited via the hub schools' newsletters and events. The developers felt that this recruitment method sped up recruitment and facilitated delivery (as it limited teachers' travel time). However, they recognised that it also limited the generalisability of the evaluation findings.

There were no explicit inclusion/exclusion criteria. One of the hub schools is known to perform very well in GCSE science so it is possible that nearby schools accepted to take part in the trial in the hope of benefitting from the hub school's experience. A few schools refused to take part, but for reasons unrelated to the intervention (for example, new head of science freshly recruited and reluctant to embark on a pilot at an early stage).

SSTT was aimed to be delivered to all Year 10 students (2,100 students). There were no explicit inclusion/exclusion criteria.

As this was an intervention carried out within schools by school staff, it was not necessary to gain parental or student permission. Participation in research, however, was not compulsory. We told teachers and students that they could exercise their right to withdraw from the evaluation.

2.3. Data collection

In the interests of transparency and replicability, this section describes both the original data collection plan (see 2.3.1) and the changes made to this plan after the publication of the study plan (see 2.3.2).

2.3.1. Original data collection plan

The original data collection plan included 11 sources of data (SD), which are presented below.

SD1 included the organisation of two logic model workshops. The aim of the pre-intervention workshop was to clarify the developers' assumptions (see Appendix I). The aim of the post-intervention workshop was to critically review the logic model based on the experience of developers and to suggest improvements (see Appendix II).

SD2 included two rounds of semi-structured interviews with the heads of science in each sampled school. Early implementation interviews were conducted after the first training workshop. Their aim was to explore the motivations for participating in the pilot, experiences of the training, plans for implementation, expectations for the intervention, and how it compares to 'business as usual' for the setting. Post-intervention interviews were conducted at the end of the school year. Their aim was to explore the feasibility and acceptability of the SSTT, barriers and facilitators to delivery, perceived impact for professional practice, and student outcomes. These interviews were also intended to explore recommendations for improvement to the SSTT and to the support interventions (for example, training, parent engagement, etc.). These interviews were conducted on the phone. Interviews were recorded and transcribed by an external agency.

[3] https://educationendowmentfoundation.org.uk/public/files/Projects/SSTT_Evaluation_Plan_FINAL.pdf

SD3 included two rounds of semi-structured interviews with sampled teachers. Early implementation interviews were meant to be conducted just after the first coaching session. Their aim was to assess the acceptability of the SSTT among teachers (both for students and for themselves), explore teachers' strategies to boost retention and understanding (pre-intervention), and evaluate the quality of the coaching session. Post-intervention interviews were conducted at the end of the school year; their aim was to assess the perceived effect of the intervention on students' retention. These interviews were conducted on the phone. Interviews were recorded and transcribed by an external agency.

SD4 included the observation of all training workshops (that is, both day one and day two in each hub). Their aim was to assess how information was to be cascaded from hubs to heads of science. Observations were conducted by NatCen researchers using an observation template.

SD5 included the observation of one in-school coaching session per sampled school. Their aim was to assess how information was to be cascaded from heads of science to other teachers.

SD6 included the observation of one science class per sampled teacher. Their aim was to understand (1) how information was to be cascaded from teachers to students, (2) the dosage and fidelity of implementation at class level, and (3) teachers' and students' engagement with the toolkit.

SD7 included the observation of three parental information sessions (one per hub). Their aim was to assess (1) how information was to be cascaded from schools to parents, (2) parents' understanding of the benefits of self-testing, and (3) the acceptability of the toolkit.

SD8 included an online survey administered to all participating students. The aim of the survey was (1) to assess the perceived effect of the toolkit on learning and well-being outcomes and (2) to estimate the response rate, should the intervention be trialled. The survey questionnaire was developed by the evaluation team and was divided into four categories: (1) background, (2) attitudes to learning and memorisation, (3) learning strategies before the intervention, and (4) experience of using the SSTT.

In the absence of named sample, the survey was administered to all students through their head of science. Data was captured using the Build software.

The survey was designed to take 15 minutes to complete. The survey questionnaire can be found in Appendix III.

SD9 included any app and/or usage data that would be generated during the project from any digital tool (SSTT, parent toolkit).

SD10 included cost data to assess the affordability of the intervention. We provided developers with a pro-forma to estimate the costs of developing the SSTT and delivering training and support to schools.

SD11 included secondary data collected from the DfE School Database about participating schools. The aim was to describe the profile of pilot schools based on key characteristics. We compared pilot schools with the average school in the U.K. based on OFSTED rating, GCSE results, class size, proportion of FSM students, and proportion of BME students and so forth.

2.3.2. Post-design changes to data collection

Some research activities were cancelled. The evaluation team cancelled all pre-intervention teacher interviews and a significant number of observations due to the unresponsiveness of some schools and teachers. Teachers cancelled other research activities. For example, one school found that the scheduled class observation was too close to the end of term, which is a busy period for school staff.

Three schools out of 12 did not complete the student survey. This suggests that the head of science did not forward the link to students or did not encourage them to take part.

One interview recording was lost due to a technical problem with the recording software.

The table below shows the scope of the evaluation, as it happened.

Table 1: Post-design changes to the data collection plan

Source of data	Description	Implementation
SD1	Logic model workshops	As planned.
SD2	Heads of science interviews	As planned.
SD3	Teacher interviews	All pre-intervention interviews cancelled by the evaluator (to reduce the burden on schools); 5/6 post-intervention interviews conducted; 1/6 post-intervention interview cancelled by the school.
SD4	Training workshop observations	5/6 observations conducted; 1/6 observation cancelled by the evaluator.
SD5	Coaching observations	1/6 observation conducted; 5/6 observations cancelled by the evaluator (to reduce the burden on school).
SD6	Science class observations	5/6 observations conducted; 1/6 observation cancelled by the school.
SD7	Parental information session observations	All observations cancelled by the evaluator (to reduce the burden on schools).
SD8	Student survey	As planned.
SD9	App/usage data	Cancelled (we wrongly assumed that some tools would be shared through websites and apps that would generate usage metrics).
SD10	Cost data	As planned.
SD11	Desk research	As planned.

These changes had the following effects on our findings:

- SD5: limited evidence of the amount and quality of support provided by the developers during the pilot; limited evidence of the amount and type of questions asked by teachers;
- SD7: very limited evidence of parents’ interest, support, and role in the project; and
- SD9: no direct evidence that all teachers downloaded and used the materials created by the developers.

2.4. Sampling

2.4.1. Original sampling plan

Table 2: Sampling plan

Unit of analysis	Number	Sampled	Rationale
Organisations			
School hubs	3	3	All school hubs were supposed to be included in the evaluation.
Participating schools	12	6	We aimed to include two schools in each of the three hubs. These schools were to be purposively selected to provide range and variation, including with regard to: size, existing practice with regard to ST, and GSCE results.
		6	SD9: we aimed to collect monitoring data from each of the six sampled schools.
		6	SD10: we aimed to provide heads of science in each of the six sampled schools with a pro-forma to help them assess the cost of the intervention at school level.
		12	SD11: we aimed to conduct desk research to describe the profile of pilot schools based on key characteristics.
People			
Developers	3	3	SD1: The logic model workshops aimed to involve one developer from each of the three Hubs.
Heads of science	12	6	SD2: we aimed to interview the head of science in each of the six sampled schools.
Teachers	36	6	SD3: we aimed to interview one teacher in each of the six sampled schools. These teachers were meant to be selected with a view to provide a range of views regarding the acceptability of self-testing and current teaching practice.
Students	2,100	All	SD8: we aimed to survey all participating students across all participating schools. The aim was to maximise sample size.
Parents	Unknown	Unknown	SD7: we aimed to gather feedback from parents participating in the information sessions in each sampled school.
Events			
Training sessions	6[4]	6	SD4: we aimed to observe all training sessions (that is, two sessions in each of the three Hubs).
In-school coaching sessions	24[5]	6	SD5: we aimed to observe one coaching session in each of the six sampled schools. The evaluators aimed to attend the first coaching session in half of the sampled schools and the second coaching session in the other half of these schools.
Science classes	Unknown	6	SD6: we aimed to observe one science class for each of the six sampled teachers.
Parental information sessions	12[6]	3	SD7: we aimed to observe one parental information session in three of the six sampled schools (and one in each of the three regions). The objective was to seek to obtain a range of locations (urban/rural) and school performances based on information provided by heads of science and our own research.

[4] 2 training days x 3 school hubs
[5] 2 coaching sessions x 12 participating schools
[6] 1 information session x 12 participating schools

2.4.2. Post-design changes to the sampling plan

Post-design changes to the sampling plan concerned the selection of schools for site visits and the selection of teachers for interviews. Given the small number of schools to be sampled, we chose only one sampling criterion, namely the size of the school. We selected the largest and the smallest school in each hub. As one school dropped out of the evaluation, we replaced it with the school of similar size. We also ensured that the sample of schools did not include more than one hub school. Overall, the achieved sample was close to the planned sample.

We selected one teacher per sampled school. These six teachers were selected in collaboration with the heads of science. Heads of science received a questionnaire asking, for each science teacher, (a) the number of years of teaching experience and (b) whether they had a degree in a science subject. We collected data for 12 teachers from six schools. Our aim was to select teachers to maximise variation in terms of background and length of experience. However, three teachers declined to take part in the evaluation and had to be replaced. In light of the above, our assessment is that the achieved sample is closer to a convenience sample than originally planned. It is likely that the views expressed by teachers in this report are somewhat more favourable to the SSTT than those of the 'average' teacher.

2.5. Analysis

2.5.1. Original analysis plan

Qualitative data

The study plan specified that we would analyse raw, qualitative data thematically using the Framework approach. The aim was to analyse the data by theme and by case. We also considered performing within-case analysis to triangulate the perspectives of providers and developers across the year in order to come to a holistic picture of pilot implementation.

Quantitative data

The study plan specified that raw survey data would be analysed by means of frequencies and cross-tabulations but did not indicate which variables would be cross-tabulated and how.

2.5.2. Post-design changes to the analysis plan

Qualitative data

Interviews were digitally recorded and professionally transcribed. In line with the study plan, we used Framework to chart and analyse transcribed data. We employed a mixed deductive/inductive approach for the charting of transcripts and observation notes with data being synthesised according to both pre-established themes as set out in the topic guides and templates as well as emerging themes. For each research question, we compared the expectations or assumptions of the developers (using information gathered during the logic model workshops) with our own observations and notes to assess the extent to which the intervention was delivered as planned and met its objectives.

Given the new method for sampling teachers (largely based on convenience), the reduced number of interviews, and the low level of response variation, we decided not to do within-case analysis. It was not possible to contrast the views of more senior and more junior teachers as initially planned. As a result, the evaluation has slightly less depth than initially planned.

Quantitative data

At the analysis stage, we selected two background variables that were expected to be correlated with the outcomes of interest, namely gender and science self-efficacy. The latter variable was a scale derived from NatCen's bespoke student survey questionnaire.

The possible answers to the gender variable were 'male', 'female', 'other', and 'prefer not to say'. We recoded 'other' and 'prefer not to say' as missing values due to small frequencies.

The survey also asked students five questions relating to self-efficacy. It asked (1) how they feel about the effort they put in science, (2) the marks they get, (3) their enjoyment of science, (4) how easy they found memorising and understanding science lessons, and (5) how confident they feel about memorising for tests and exams.[7] Possible answers were on a five-point Likert scale ranging from 'strongly agree' to 'strongly disagree'. Using these five questions, we created a science self-efficacy scale ranging from five to twenty-five, with lower scores indicating greater self-efficacy. We found the scale to be reliable, with a Cronbach's Alpha coefficient of 0.8. For our analysis, we calculated the mean self-efficacy score across the five items. To simplify interpretation, we then turned this score into a dichotomous variable: students scoring below the mean were put into one group and those scoring on the average or above in another. We found that 53% of respondents were in the higher self-efficacy group and 47% in the lower self-efficacy group—a result close to what we expected.[8]

We then cross-tabulated gender and self-efficacy with the student survey. We tested the significance of the association between two variables using chi-square tests. We used p-values to assess the degree of compatibility of our statistics with our initial hypothesis. We conducted all analyses in SPSS 25 and Stata SE16. We reported all data manipulations, analyses and results..

Cost data

Cost data was analysed using the EEF methodology (EEF, 2015).

2.6. Timeline

Table 3 presents key dates and research activities for the evaluation.

Table 3: Project timeline

Month	Research tasks
September 2018	Set-up meeting.
December 2018	Pre-intervention logic model workshop.
January 2019	First interview with heads of science; observation of training workshop 1.
February 2019	Observation of coaching session 1; observation of training workshop 2.
June–July 2020	Second interview with heads of science; interview with teachers; observation of science classes; student survey.
October 2020	Post-intervention logic model workshop; cost data collection.
November 2020	Desk research; qualitative data management and analysis; cost data analysis.
January 2020	Draft report submitted.
February 2020	Peer review.
March 2020	Submission of data to EEF archive.

[7] See Appendix III for the detailed phrasing of the question.
[8] The slightly higher proportion of high-efficacy students reflects the fact that 'average' students were put in the high efficacy group.

2.7. Reporting of findings

Our initial plan was that the findings of the evaluation would be reported following the three IPE dimensions mentioned in section 1.3. However, we modified this structure at the drafting stage to improve readability. The table below outlines the structure of the Findings chapter and whether/how each dimension in this report relates to the dimensions in the study plan.

Table 4: Structure of the Findings chapter

Section #	IPE dimension (in the report)	IPE dimension (in the study plan)	Reason for change
3.1.	Profile of participants	Scalability	We felt that, given the small-scale and qualitative nature of the pilot, one could not expect participating schools/students to be ‘representative’ of the population of schools/students in England.
3.2.	Usual practice	--	Added as a separate section to reduce the length of the Feasibility section and put greater emphasis on how SSTT differs from business as usual.
3.3.	Evidence of promise	(As planned)	N/A
3.4.	Feasibility (toolkit)	Feasibility	The drivers/obstacles to delivering the SSTT were analysed separately from the drivers/obstacles to implementing support activities (in the interest of readability).
3.5.	Feasibility (support activities)	Feasibility	As above.
3.6.	Costs	--	Added as a sub-section to reduce the length of the Feasibility section.
3.7.	Readiness for trial	--	Added as a finding rather than as a conclusion to clearly address the question of whether the SSTT should be trialled.

3. Findings

3.1. Profile of participants

The aim of this section is to describe the profile of participants based on key characteristics likely to affect the outcomes of interest. As this was a small-scale pilot, no generalisation of the findings from the sample to the population was attempted, with the exception of the student survey.

3.1.1. Schools in the pilot

Compared with the average secondary school in England,[9] the average pilot school was larger (1,114 students in the pilot vs 983 students in the population), had a significantly lower percentage of students eligible for free school meals (6.5% in the pilot vs 15.5% in the population), was more likely to be rated 'good' or 'outstanding' by Ofsted (92% in the pilot vs 80% in the population)[10] and more likely to be in a rural area (33% in the pilot vs 15% in the population). In light of the above, NatCen's assessment is that the conditions of implementation were more favourable than what they would be in a trial. No school dropped out of the programme during the pilot.

3.1.2. Schools in the evaluation

Not all pilot schools participated in the evaluation.[11] We sampled six schools for our qualitative research. Compared with the average pilot school, the average sampled school was smaller (844 students in the sample vs 1,114 students in the pilot), had a similar percentage of students eligible for FSM (6.7% in the sample vs 6.5% in the pilot), was less likely to be rated 'good' or 'outstanding' by Ofsted (83% in sample vs the 92% in the pilot), and equally likely to be rural (33% in the sample and the pilot).

3.1.3. Participating teachers

The teachers whose views are represented in this report were sampled with the aim of providing a broad range of experiences, both quantitatively (number of years of teaching experience) and qualitatively (science subject taught or studied). This was achieved. The sampled teachers had between two and 23 years of experience and had studied a broad range of subjects, including physics, chemistry, and biology. However, it is important to note that the alterations made to the initial evaluation design are likely to have introduced some confirmation bias in our findings. The initial design included two waves of interviews (pre-intervention and post-intervention) with (a) the head of science and (b) one teacher in each participating school. As mentioned previously, all but one pre-intervention interview with teachers were cancelled due to non-response. The consequences of this alteration are twofold: first, the views reported in this evaluation are those of the most responsive teachers; second, and most importantly, the views of heads of science are overrepresented in this report. This is not a trivial fact considering that heads of science are those who volunteered to take part in the pilot and coordinated its implementation in schools.

3.1.4. Participating students

Given the research questions listed in section 1.3, the early development stage of the intervention, and the limited scope of the project (12 schools), NatCen did not collect data about participating students other than the data volunteered by students through the online survey.

Of the 2,100 students who took part in the pilot, 1,040 from nine schools completed the survey. Three cases from three schools were excluded as the very low response rate suggested that the survey had not been disseminated as requested. Therefore, the findings below are based on the views of 50% of participants in 9 out of 12 schools.

---

[9] Based on figures collected from the DfE school database.
[10] Based only on schools for which the rating is available on the DfE school database.
[11] The pupil survey is the only data collection tool that aimed to involve all schools.

In the absence of a sample frame, it was not possible to compare respondents and non-respondents. However, the distribution of two background variables suggests that the sample was reasonably close to what would have been expected with a higher response rate:

- Gender: there were more males (51%) than females (42%), which could be partly attributed to the participation of a school for boys and the non-participation of a school for girls. Around 8% of respondents chose the option ‘other’ or ‘prefer not to say’.
- Science self-efficacy: 53% of the sample scored average or above on the self-efficacy scale, while 47% scored below average. This slight imbalance can be explained by the fact that respondents who neither agreed nor disagreed with each statement forming the scale were arbitrarily assigned to the higher self-efficacy group to create a binary variable and simplify cross-tabulations. The construction of the self-efficacy scale is explained in section 2.5.2.

3.2. Usual practice

To understand students’ and teachers’ assessment of the SSTT, it is important to understand what would have happened in the absence of the SSTT (usual practice).

3.2.1. Developers’ assumptions (from the logic model workshop)

During the pre-intervention workshop, the developers acknowledged that some of the tools in the SSTT were likely to already be used by teachers (for example, flash cards, cumulative quizzing). However, they considered the novelty of the SSTT to be in the self-testing approach—for example, teachers using flash cards were unlikely to use them in a self-testing way. Developers were interested to find out whether the pilot would change the way teachers use some of these strategies. Other strategies in the SSTT were expected to be more novel, like mind-mapping tests and structured note-taking.

The developers also acknowledged that some teachers might have used strategies like flash cards and cumulative quizzing in Year 9. Thus, some students were expected to be familiar with some of these strategies, but not inclusive of the self-testing element.

3.2.2. Usual practice according to teachers

The developers’ assumptions were supported by interviews with teachers and heads of science. It is apparent from these interviews that schools had previously used the strategies named in the SSTT. All teachers who were interviewed reported using at least one SSTT strategy for teaching during lessons or for setting as homework. However, there was a strong sense that when using these strategies, teachers did not usually include the self-testing element. Teachers who did include the self-testing element within their usual practice explained they had not used these strategies systematically. In support, heads of science noted that the use of the toolkit strategies was inconsistent amongst teachers. Therefore, the combination of five clearly defined strategies which incorporated a self-testing element was viewed as novel by heads of science and teachers.

Interviews with heads of science reveal that they were not only familiar with some of the strategies in the SSTT, they were also cognisant of, and conversant with, the concepts of self-testing and meta-cognition, which underpin the SSTT. Similarly, heads of science felt most teachers had some awareness of self-testing, particularly those who are reasonably well informed about research or who regularly attended training.

3.2.3. Usual practice according to students

The student survey results confirm that some strategies had already been used by participants. In fact, the toolkit includes the two preferred learning strategies of respondents pre-intervention, namely flash cards (used by 65% of respondents) and mind maps (used by 39% of respondents). The three other strategies in the toolkit were either new to most students or known but untested. Pre/post tests, structured note-taking, and cumulative quizzing were only used by between 6% and 7% of respondents pre-intervention.

Respondents found pre-intervention strategies most helpful for short-term memorisation (56% of respondents found them helpful or very helpful for memorising a lesson for an exam). Less than half of respondents (47%) found them helpful or very helpful for understanding ideas and concepts of science lessons. About a third (35%) of students said their prior strategies were helpful or very helpful for long-term memorisation. Looking at negative perceptions, over half (55%) of students indicated that their prior strategies were either ‘very’ or ‘quite’ time-consuming to get through materials.

3.3. Evidence of promise

Evidence of promise was defined as the extent to which the intervention delivered its main outcomes—students’ understanding and retention of science lessons.

3.3.1. Developers’ assumptions (from the logic model workshop)

The developers expected that the toolkit would support students’ short-term and long-term retrieval of factual information. The developers acknowledged that this may not lead to better attainment directly, but instead predicted that if students were better able to retrieve facts during exams, they would have lower cognitive demands and, therefore, be more able to apply their knowledge to exam answers, particularly for questions requiring lots of content.

The developers predicted that the effect of the SSTT would be higher for:

- students who were less familiar with each of the five strategies before the pilot; and
- students with high attendance due to more opportunities to engage with the toolkit.

The developers predicted other subgroup effects, but were unsure about the direction of these:

- Disadvantaged students. Although the developers recognised that disadvantaged students may have the most to gain, they may also lack a supportive home environment and therefore not make the expected gains.
- Middle-attaining students. The developers predicted middle-attaining students would have the most to gain from the SSTT as they typically lack strategies for long-term memory but are confident and already have some successful learning strategies in place.
- High-attaining students. The developers felt high-attaining students may be most reluctant to change their learning strategies because they have already achieved well in the past with their existing strategies. However, the developers predicted that if they are willing to change their approach to revision strategies by incorporating a self-testing element, they may develop more efficient strategies and therefore benefit from the SSTT.

3.3.2. Evidence of promise according to teachers

Heads of science were all convinced that, on average, students would benefit from the intervention and those with prior knowledge of the evidence base around self-testing viewed the toolkit positively as they were aware of the potential benefits of using the strategies. For example, they predicted the strategies would encourage students to revise independently and improve their retention and recall in science, which, in turn, would result in better revision and exam performance. Alongside improved attainment in science, heads of science recognised that the toolkit had potential for whole-school improvement due to transferable skills from science to other subjects.

Heads of science felt the toolkit would be particularly beneficial for science compared to other subjects. They acknowledged the subject follows a linear structure, meaning all content is tested at the end of GCSEs. They also commented that GCSEs are now more content-heavy so felt the toolkit would be beneficial for students and teachers to help overcome the barriers associated with learning large amounts of information for exams. They also recognised that as science is a compulsory rather than an optional subject, students sometimes lack motivation for science and so would benefit from a targeted intervention to improve recall and retention.

Heads of science were also aware of the limitations of the SSTT. They perceived that the narrow focus of the SSTT as its main limitation. heads of science were conscious that recalling is not enough to do well in a science exam; other skills are needed, such as pure maths. They were concerned students would view the toolkit as a 'quick fix' for achieving well in science exams and would disregard the importance other skills, such as applying their knowledge to different contexts.

> *'But I just wonder if lessons will be too focused on recalling information and not enough on applying it and working scientifically skills' (head of science #6, pre-intervention).*

Another concern raised by some heads of science was related to how students of different abilities would respond to the toolkit. In particular, they were concerned about the use of self-testing with lower attaining students who may have less confidence in their science ability. This led to the view that the toolkit may widen pre-existing educational gaps between high and low attaining students.

> '*What we notice is quite often with those learners who are high-attaining, they have a lot of these skills independently and they're very keen to identify key ideas and quiz themselves and make the resources. Whereas with those students who make less progress and who are lower attaining, they are much more reluctant to engage with this sort of activity and will find it more difficult' (head of science #2, pre-intervention).*

3.3.3. Evidence of promise according to students

3.3.3.1. Experience of using the SSTT

In this section we report descriptive statistics relating to the experience of using the SSTT (see Appendix IV for detailed results). Overall, almost two thirds (62%) of students said they want to keep using the SSTT in class and a similar number of respondents (64%) said it will be helpful to understand and revise other subjects.

The most effective strategies for **remembering** science lessons were flash cards (for 46% of respondents), structured note-taking (15%), mind-mapping tests (12%), pre/post tests (8%), and cumulative quizzing (7%). One-in-twenty (5%) respondents reported that no strategy was useful for remembering science lessons.

The most effective strategies for **understanding** science lessons were flash cards (for 37% of respondents), structured note-taking (24%), mind-mapping tests (11%), pre/post tests (10%), and cumulative quizzing (8%). Around 7% of respondents reported that no strategy was useful for understanding science lessons.

The most **enjoyable** strategies were flash cards (for 50% of respondents), structured note-taking (12%), mind-mapping tests (12%), pre/post tests (6%), and cumulative quizzing (6%). One in ten (10%) students reported not enjoying any strategy.

The fact the order of responses is the same for the three outcomes of interest (remembering, understanding, enjoying) confirms that the three outcomes are closely interwoven.

Some of the results mirror the responses to the 'usual practice' question. More specifically, the strategy found to be most effective and enjoyable (flash cards) was also the most used strategy pre-intervention. Other results are more surprising. For example, structured note-taking was one of the least used strategies pre-intervention but ended up being the second most promising strategy post-intervention. It should also be noted that structured note-taking was found more useful for understanding science lessons (24%) than for remembering them (15%).

The questionnaire also included a question about a possible downside of self-testing, namely stress. A relative majority of respondents (40%) found using the toolkit 'a little bit stressful', 30% of students found the toolkit 'very' or 'fairly' stressful and only one-in-five found it 'not stressful at all'.

3.3.3.2. Association between science self-efficacy and experience of using the SSTT

Exploratory analyses were conducted using cross-tabulations and chi-square tests to find out whether students with higher science self-efficacy had more positive experiences of the SSTT than students with lower science self-efficacy (see Appendix V for detailed results).

A positive association was found between self-efficacy and frequency of use. Indeed, almost half (46%) of those with a higher self-efficacy score reported using the SSTT more than once or twice a week, whilst three in ten (29%) of those with lower scores of self-efficacy reported so (p-value < 0.001).

Similarly, a positive association was found between self-efficacy and respondents' assessment of the toolkit. Seven out of ten (70%) of these students found the tool helpful or very helpful to remember a science lesson for a test, compared to only 42% for students whose self-efficacy was below average (p-value < 0.001). Similar findings were found for the effectiveness of SSTT in improving understanding and remembering a science lesson in the long-term.

These results mirror students' assessment of their pre-interventions strategies. Indeed, almost two thirds (62%) of students with higher self-efficacy indicated that their previous strategies were helpful for remembering a science lesson for a test or exam, while less than a third (28%) scoring below average indicated so (p-value < 0.001). Similar percentages were found when asked about how effective prior strategies were to understand science lessons and remember them in the long-term: students with lower self-efficacy were more likely to report that their previous strategies were very time consuming (18% vs 10% for those with higher self-efficacy).

There was no marked difference between self-efficacy groups in what they found most enjoyable or most helpful to understand science lesson: both groups thought the most enjoyable and helpful strategy was flash cards (p-value = 0.328). Nevertheless, students in the low self-efficacy group were more likely to find none of the SSTT strategies helpful in remembering science lesson (9%) compared to those in the higher self-efficacy group (3%; p-value = 0.032).

Students in the higher self-efficacy group were less likely to find it stressful: seven out of ten (71%) of students in the higher self-efficacy group found it a little bit stressful to not stressful at all compared to 53% in the lower group (p-value < 0.001). This is underpinned by the fact that a larger proportion of students in the higher self-efficacy group said they would want to continue using SSTT in class (74% vs 50% of those in the lower self-efficacy group, p-value < 0.001). Similar proportions were found when students were asked whether they think the toolkit strategies would be helpful to understand and revise other subjects.

3.3.3.3. Association between gender and experience of using the SSTT

Exploratory analyses to assess gender differences were also conducted, although the expected direction of the effect was unclear (see Appendix VI for detailed results). There was no evidence of an association between gender and how effective the toolkit is perceived to be for helping understanding and remembering science lessons in the short term or long term. There was also no evidence of an association between genders and the frequency at which the toolkit is used. Nevertheless, females were less likely to enjoy pre-test and post-test strategies (2% vs 9% of males); a higher proportion of them preferred flash cards (57% vs 46% of males; p-value < 0.001).

There was also strong evidence of an association between gender and how stressful the toolkit was perceived to be. A quarter (25%) of males reported finding the toolkit not stressful at all while only 17% of female said so (p-value = 0.004). Nevertheless, there was no strong evidence of a difference between females and males when it came to whether they wanted to continue using the toolkit in class or whether they thought the strategies would be helpful to understand and revise other subjects.

3.4. Feasibility of integrating the intervention into the curriculum

'Feasibility' was defined as the sum of all drivers and obstacles to the success of the intervention. This section focuses on the feasibility of integrating the SSTT into the Year 10 science curriculum. The feasibility of delivering support activities to all participating teachers is discussed separately (see 3.5).

3.4.1. Developers' assumptions (from the logic model workshop)

In terms of behaviours, the developers had four main expectations about the delivery of SSTT. They expected teachers to:

- introduce the SSTT to students immediately after the first training workshop;
- use at least one strategy from the toolkit in every science lesson;
- try all strategies in the toolkit; and
- make some adjustments to the toolkit to reflect local practices and needs.

In terms of attitudes, the developers expected that most teachers would welcome the SSTT. However, they also predicted that attitudes towards the SSTT would be more positive among:

- less experienced teachers—they may be more willing to use the toolkit due to being more open-minded, and/or less confident in their teaching ability; and
- teachers who attain below-average results—they may feel their current strategies are ineffective and be more eager to use different teaching strategies.

3.4.2. Feasibility of integrating the SSTT into the curriculum according to teachers

3.4.2.1. Introduction of the SSTT

Interviews with heads of science and teachers suggest a high level of compliance. Several teachers used the slides provided by the developers to introduce the toolkit to students. The slides were found most helpful to explain strategies with which students were thought to be less familiar (for example, structured note-taking). Teachers also summarised the research underpinning the toolkit to maximise engagement to the students.

In contrast, some teachers did not use the slides provided by the developers to introduce the toolkit to students. They were worried that students would not respond well to the formal structure of the toolkit so wanted to keep the toolkit light touch when introducing it to students.

> *'Rather than making a big scene, this is, because it looks, it is a really easy-to-use document, but for 15-year-olds, looking at that, it's quite … it looks quite formal! Although it's really great, but for a 15-year-old, it'll look like a proper document, like how to learn, I think they'd see it, and I think that may put them off' (science teacher #3, post-intervention).*

Importantly, some teachers indicated that they had introduced the SSTT to Year 9 students thereby exceeding the developers' expectations.

3.4.2.2. Frequency of use

The frequency at which the SSTT was used in class is unclear. The evidence collected through interviews and observations suggest the SSTT has been used quite extensively. Teachers reported using the strategies in most lessons.

> *'Well, I'm trying to get it in as a regular part of the lessons, but I think there's certainly some where it's harder to find a way to fit it in than others, but on the whole, yes, it's as many lessons as possible' (science teacher #4, post-intervention).*

However, additional evidence suggests the toolkit was not used in every lesson. For example, a relative majority of the students who completed the survey recalled using the toolkit once every few classes (see section 3.4.3). Furthermore, out of the five classes observed by the evaluation team to assess the use of the SSTT, one did not use the SSTT.

Some heads of science reported that teachers in their school had not only used SSTT in class but also incorporated a self-testing approach when setting homework.

3.4.2.3. Duration of use

The expected duration of use of the SSTT in class was not discussed in the logic model workshop. We assessed this duration by means of direct observations of science classes. We found great variation between teachers. In one observation, the SSTT was not used at all. In other observations, we found that SSTT was extensively used. For example, in one of the classes, the SSTT was used for the entire duration of the lesson (that is, 55 minutes out of the hour lesson, accounting for the time it took to set up and pack away the lesson). In this lesson the students used a combination of structured note-taking and cumulative quizzing which was led by the teacher. In another class, the teacher introduced topics at the start of the lesson and the students spent the remaining time (approximately 40 minutes) completing structured note-taking and then using the questions to self-test each other in pairs.

3.4.2.4. Using a mix of strategies

Developers expected that, in the long run, teachers would become adept at choosing the strategy that would be most suited to the lesson's intended learning outcomes. In line with this expectation, teachers demonstrated an awareness around the utility of the different strategies. For example, pre/post tests and cumulative quizzing were thought to be better suited for the start and end of topics and mind-mapping tests to be better suited to homework.

The evidence suggests that teachers did try all strategies but used some at a much higher frequency than others. Heads of science mentioned that teachers use flash cards and cumulative quizzing extensively because they are quite easy to implement and are very accessible. Regarding structured note-taking, teachers stated that it is easier to implement this strategy in top set science classes as these students tend to do it well and find it more useful.

3.4.2.5. Customisation

Most teachers who were interviewed used the toolkit but made their own individual adaptations to allow the toolkit to fit their students' learning needs.

Some teachers made changes such as reducing the amount of text that students need to write for structured note-taking. Methods to provide more scaffolding included providing a template to help students structure their notes and adapting the revision guide to leave space for students to write their own questions and notes in the revision guide rather than their exercise books.

In line with concerns about mind-mapping, teachers provided a skeleton outline or a standard format to assist students with this strategy.

> *'So whatever the content, we get a clock starting in the top left corner and working round the mind map, so that our students have this constant format to follow to try and improve their memory and retrieval. So we shared that' (science teacher #5, post-intervention).*

Another common adaptation was to incorporate technology, such as websites or apps, within the strategies in order to, for example, create quizzes.

Some strategies were tweaked or changed to suit students of different abilities. Teachers felt it was necessary to provide students of lower ability with more support when using the toolkit. This included providing questions rather than asking them to write their own as well as checking whether the answers written down were correct.

3.4.2.6. Perceived effect of SSTT on teaching practice

Teachers reported that the SSTT changed their teaching practice in a number of ways. The focus on self-testing across the five strategies has encouraged teachers to place more emphasis on students' learning and remembering content taught in lessons. In addition, teachers reported the toolkit helped them advise students on how to revise and self-assess. Furthermore, the five clearly defined tools helped teachers promote a consistent approach to self-testing amongst students.

Teachers also reflected on how their usual classroom practice had changed in relation to specific tools. For example, teachers reported their use of flash cards had changed. In line with the toolkit, they encouraged students to self-test by writing questions on one side and answers on the other whereas previously students were writing blocks of text on each side of the flashcard.

3.4.2.7. Factors facilitating the integration of the SSTT into the science curriculum

Interviews with heads of science and science teachers indicate that staff had a rather high level of engagement with the intervention. In general, teachers found it easy to adapt lessons and include strategies into their usual classroom practice. Furthermore, some teachers felt the toolkit has given lessons more structure and has allowed for links to be made to exams, such as through exam-style questions being used within each of the tools. In line with the positive views about the toolkit, all heads of science indicated that they were planning to continue using the toolkit beyond the pilot.

Key factors of the toolkit which facilitated its integration into the Year 10 curriculum included:

- Linear GCSE structure—heads of science acknowledged the linear structure of GCSEs requires students to memorise large amounts of content, and so teachers are open to incorporating strategies which aim to improve memory retention and recall.

'The new GCSEs and the emphasis on content and recall, I think teachers know that there is a need to address helping students to revise. So, yes, it was embraced really positively. There wasn't any kind of resistance to it' (head of science #3, post-intervention).

- Evidence base—the evidence base for each of the tools helped to persuade heads of science that it would be beneficial to use the SSTT in their school. The evidence base contributed to high buy-in amongst teachers who viewed the SSTT as low risk.
'It's based on research, so I know it's going to work' (science teacher #6, post-intervention).

- Involvement in research—many heads of science were motivated to sign up and take part in the pilot because of their interest in the research results. They hoped the results from the pilot would show that the SSTT is effective for improving retention and recall as they were planning to use the findings to make it easier to 'sell' the SSTT to students the following year.

'That makes the buy-in better for next year's students if I can show the efficacy of this. Do you see what I mean? It's going to be an easier sell' (head of science #5, pre-intervention).

- Graphics—the visual representation of each of the tools was thought to make the toolkit accessible to students, parents, and teachers, and particularly for low-ability students.
- Accessibility—the technical information about each of the tools was presented in an accessible way.

'Students could access it, parents could access it, teachers could access it. It was quite one size fits all, really. It was technical enough that teachers … It informed them enough to be able to make tweaks to their practice, but it wasn't so technical or so overwhelming that students or parents couldn't engage with it' (head of science #6, post-intervention)
.

- Encourages independent revision—the tools are designed to be used by students independently, which heads of science felt was advantageous as it would encourage students to be more proactive and responsible for their own revision.
- Encourages active revision—the toolkit explains why active retrieval is important, which heads of science felt would help students recognise the importance of active revision. This was thought to encourage them to use effective strategies rather than relying on passive techniques.
- Minimal impact on teachers' workload—as the tools were relatively simple and easy to understand, heads of science felt they would not greatly add to teachers' workload.
- Quick, positive results—teachers liked the fact that they could see an immediate impact on students learning and this contributed to their sustained buy-in.
- Familiarity with tools—as most of the tools were familiar, heads of science were confident in their implementation.

> *'The school had already used flash cards and cumulative quizzing so the tools didn't need a lot of explanation. We'd already done quite a lot of work on flash cards so, really, the toolkit was just a package of some things that we were already using in the department. So it didn't really need lots of explanation and lots of … it wasn't really open to that much interpretation' (head of science #3, post-intervention).*

3.4.2.8. Factors hindering the integration of the SSTT into the science curriculum

Key factors included:

- Number of tools—heads of science felt it would be difficult to introduce all five tools concurrently and worried this would be overwhelming to students. They felt it would be better to introduce the tools more gradually so that students were given enough time to familiarise themselves and embed the use of each of the tools in turn.
- Time—some teachers struggled to find time to create resources and train students effectively. Furthermore, finding the time to get staff to sit down to receive training in the strategies has been a difficult challenge in some schools. To add to this, teachers have found it difficult to find time to create new slides that are in line with the strategies meaning that they have had to use slides from previous years.
- A lack of understanding of the link between short-term memorisation and deep learning—a science teacher understood how self-testing could improve short-term memorisation, but did not equate short-memorisation and *learning*:

> *'It was revision, so it can be dry at times, but I certainly wouldn't think about teaching like that if I was introducing some new content and wanted to get them engaged with a topic' (science teacher #3, post-intervention).*

- Timeliness of the pilot—some teachers argued that it would be more effective to introduce the SSTT in Year 9 where more time can be spent on teaching new strategies. Those who did introduce the SSTT in Year 9 found that it had worked well.

3.4.3. Feasibility of integrating the SSTT into the curriculum according to students

The student survey only focused on expectations two (frequency of use) and three (using a mix of strategies). Regarding the frequency of use, the results confirm that the dosage was lower than expected. Four in ten students (42%) reported using SSTT once every few classes, whilst 28% of students reported using SSTT either in every class or every other class. Over a quarter (26%) of students said they used SSTT outside of class once or twice a week whilst 10% of them use it three to five times a week and one in five (19%) said they used the tool once or twice a month. Only a few (3.5%) did not recall using any of these strategies.

Regarding the use of a mix of strategies, the results from the survey are unclear. On the one hand, only a small minority of respondents (10%) recalled using only one strategy. On the other hand, respondents recalled using an average of two strategies only. The most-used strategy during the pilot (flash cards, at 77%) was also the most-used strategy pre-intervention (65%). The least-used strategy was cumulative quizzing (29%). The fact that respondents reported using only one strategy at home on average suggests that students stuck to their preferred strategy. Overall, the correlation between the strategy used in class and the strategy used at home was low (c. 0.2).

3.5. Feasibility of delivering support activities

This section focuses on the feasibility of delivering support activities, which included two training workshops, cascading activities, in-school coaching, twilight sessions, and parent engagement sessions. Each support activity is discussed separately below.

3.5.1. Training workshops

3.5.1.1. Developers' assumptions (from the logic model workshop)

The developers had two main expectations regarding the delivery of training workshops.

- Aim and format: it was planned that the content of the toolkit would be split across two workshops. The first workshop would train participants in using three tools while the second workshop would focus on the remaining two tools.
- Attendance: developers' expectation was that the training would be attended by two teachers per participating school, including the head of science.

3.5.1.2. Evidence of implementation fidelity

The developers made minor adjustments to the training workshops prior to implementation: they adapted the programme so that Workshop 1 would cover all tools. This is because the developers felt the strategies are not complex or unfamiliar and so it would be possible to teach all five strategies within the first workshop. Workshop 2 was used to revise the tools and prepare for the parental information session.

Both the head of science and a science teacher from each school attended Workshop 1 and 2.

3.5.1.3. Participants' experiences of the training workshops

Heads of science were motivated to take part in the training as they felt the use of clearly specified tools would make their approach to teaching more systematic and consistent. They were keen to learn techniques that could be used with students in a constructive and effective way. The support activities were viewed as particularly appealing for newly qualified teachers.

> *'It's also about engaging staff. Good quality professional development is really important for keeping staff in the school' (head of science #2, pre-intervention).*

Overall impressions of the training workshops were positive. According to heads of science, the workshops were organised and delivered well. They felt the training was comprehensive, accessible, and explained the toolkit in sufficient detail. They felt the level of theory was appropriate although noted that if they had not been previously familiar with self-testing strategies, they would have needed the theory explained in more detail. They commented that it was helpful to receive a physical copy of the EEF report on metacognition (Quigley, Muijs, Stringer, 2018) as it helped them understand the rationale behind the toolkit and the evidence base. One teacher commented:

> *'It was nice to feel that what we were going to implement actually had a secure background in terms of it was tried and it was tested, and these are the gains that potentially we could get from them, rather than just, "This is something nice I saw, let's have a try"' (science teacher #6, post-intervention).*

The training resources were thought to be of high quality. In particular, heads of science noted the slides were useful as they could be used and adapted for their own use when presenting the toolkit to teachers in their schools.

> *'I think the summaries will be really useful, I think but the slides will be—the fact that you can take it, you can use it, you can adapt it, you can [use it for] staff training yourself or you can use some of them with the students. I think it's been really well constructed so that you have the things that you will need without having to do a lot of additional work, which I think is always a big killer with these sort of things where in order to use the material that you've been given, you'd have to change it or it's not particularly user-friendly as it is now' (head of science #2, pre-intervention).*

Heads of science who attended the training found it beneficial to speak to colleagues from other schools, facilitated by the relatively small group of teachers attending the training, and would have liked more opportunities to do this within the workshops. They felt this would have allowed them to share experiences of implementing the toolkit, including modifications and methods to successfully embed the strategies.

Three concerns were expressed by heads of science about the workshops. Firstly, teachers were concerned about taking time out of class to attend training. They noted difficulties with finding cover and concerns that missing a lesson would disadvantage students. The second concern was related to the training schedule: attendees noted they would have liked more time to reflect on their learning between the first and second workshop. Finally, attendees felt it would have been beneficial to have more material and time dedicated to planning how they were going to implement the tools.

3.5.2. Cascading activities

3.5.2.1. Developers' assumptions (from the logic model workshop)

The developers had the following expectations regarding the delivery of cascading activities.

- Aim and format: as it was not possible to train all science teachers in each of the 12 pilot institutions, the developers expected that trained participants would then cascade this training to other science teachers in their school. They expected that cascading activities would cover the aims of the pilot, the science supporting the intervention, and some tips for using each of the five strategies.

- Beneficiaries: the developers expected the trained participants to cascade training to all Year 10 science teachers in the school and therefore expected all Year 10 science teachers to deliver the intervention to their Year 10 students.

3.5.2.2. Evidence of implementation fidelity

There was strong evidence of cascading. Heads of science cascaded their learning from the workshops to other science teachers in the school as planned. Across schools, initial cascade training was delivered in a single pre-planned departmental meeting by the head of science and the teacher who attended the workshops. The session usually lasted between 30 minutes and one hour. Holding the cascade training during pre-planned meetings was thought to promote attendance.

In most schools, the initial cascade training was a condensed version of the training workshops. The session was focused on presenting the existing theory and evidence base and introducing the five different tools. The session also gave teachers the opportunity to co-plan the implementation of tools within lessons.

In one school, the head of science adopted an alternative method for delivering the cascade training. Rather than introducing all five tools, the head of science instructed teachers to select one strategy based on their interest and fit with teaching style and then to train other teachers in their selected strategy. The teacher in question felt this method would reduce teachers' workload and create a sense of expertise.

> *'If they just choose one [tool] that they can engage with, then it makes them the expert and confident and they can teach other people and if people see it working for you then you're more likely to try it yourself and it snowballs that way' (head of science #5, post-intervention).*

After an initial introduction to the toolkit, heads of science arranged further follow-up training sessions, often delivered in regular weekly departmental meetings. The additional training was used to revise the strategies, discuss how students were responding to the toolkit, discuss whether they were experiencing any difficulties in delivery, and to share resources.

> *'That's why it's so beneficial to keep it as an agenda item on the department meeting to actually revisit and teachers could talk about any concerns that they had so we could overcome that as a team and come up with suggestions and ideas to move people forward' (head of science #1, post-intervention).*

Some of these follow-up sessions were informed by lesson observations and learning walks performed by heads of science. However, heads of science had difficulties finding the time to complete these monitoring activities.

Often, heads of science cascaded the training beyond the developers' expectations. Examples of initiatives taken include:

- *Cascading to other subject leads.* In some schools, the head of science and teacher who attended the training workshops met with the subject leaders (biology, chemistry, and physics) to explain the project and the tools in more detail in advance of cascading to science teachers.
- *Cascading to senior management.* In most schools, the toolkit was shared with members of the senior leadership team, usually in a brief—for example, ten-minute—informal meeting.

- *Cascading to non-science subject teachers.* In one school, the toolkit was shared amongst teachers in the school outside of the science department.
- *Cascading to local schools.* In one instance, a head of science shared the toolkit with schools in the local area during a regular meeting to disseminate good practice.

3.5.2.3. Participants' experiences of the cascading activities

Teachers saw the cascade training as crucial for successful implementation of the toolkit within the department. Overall, teachers were positive about the cascade training; they felt it was high quality and a good use of their time. In particular, teachers found it useful to be provided with the evidence base as this allowed them to explain the rationale of the strategies to students when introducing the toolkit.

Some teachers felt the cascade training could be improved by providing more resources to support successful implementation of the toolkit. In addition, some teachers would have preferred to have been provided with specific science-related examples rather than examples from other subjects.

3.5.3. In-school coaching

3.5.3.1. Developers' assumptions (from the logic model workshop)

The developers had the following expectations regarding the delivery of coaching activities:

- Aim and format: the developers had planned to deliver one coaching session in each school to support teachers in their use of the toolkit. The developers felt it would be most beneficial to deliver coaching after teachers had begun using the toolkit.

- Beneficiaries: it was intended that all teachers implementing the toolkit at each school would receive coaching.

3.5.3.2. Evidence of implementation fidelity

The developers modified the format of the coaching sessions. Rather than providing formal and structured coaching, they followed a 'question and answer sessions' approach to allow teachers to ask questions in an informal and supportive environment.

There is evidence to suggest that some teachers did not get coaching. Although the hub schools offered coaching to all participating schools, the uptake of coaching varied at the school level. Heads of science who requested coaching did so as they thought it would be helpful to receive feedback and extra support. The reasons why some heads of science did not request coaching varied between schools. While some heads of science felt it would have been helpful but did not have the time, others felt coaching was not necessary in their school.

Nevertheless, heads of science who did not receive in school coaching reported feeling well supported by their hub school and sought support through email in place of face to face support.

3.5.3.3. Participants' experiences of in-school coaching

In general, teachers found the coaching sessions very helpful and saw them as a good use of time. Specifically, it was helpful to discuss the specific problems that the school was facing as well as finding solutions to problems to ensure that the toolkit could be implemented smoothly.

> *'The staff found that really beneficial because it was tailored to what they wanted to know' (head of science #1, post-intervention).*

A number of teachers felt that anything that helps to develop teaching practice and influence students positively is a good use of their time.

3.5.4. Twilight sessions

3.5.4.1. Developers' assumptions (from the logic model workshop)

The developers had the following expectations regarding the delivery of twilight sessions.

- Aim and format: the developers aimed to deliver four twilight sessions throughout the year: briefing twilight, overview of parent resources, mid-point review twilight and celebration twilight. Each twilight was expected to be a 90-minute session after school. The developers felt the twilight sessions were an important aspect of the intervention as they gave schools the opportunity to share resources which would reduce teacher workload, and in turn, ensure sustained buy-in.

- Beneficiaries: developers' expectation was that the twilight sessions would be attended by the same two teachers who attended the workshops.

3.5.4.2. Evidence of implementation fidelity

The briefing twilight and the mid-point review twilight took place as planned. However, the following adjustments were made to the other twilight sessions during the delivery phase:

- Overview of parent resources twilight: as the resources for the parental information session were shared in Workshop 2, this twilight was cancelled.

- Celebration twilight: the name was changed to 'Next Steps Planning' as it was felt heads of science and teachers would not attend if this session was focused on a celebration. During this session, the developers shared an electronic document which schools used to outline their specific goals for sustained implementation.

3.5.4.3. Participants' experiences of the twilight sessions

In general, teachers found the twilight sessions helpful and welcomed the opportunity to share their experiences of using the toolkit with other schools. Heads of science found it valuable to have support and input from the developers at multiple time-points during implementation.

3.5.5. Parental information sessions

3.5.5.1. Developers' assumptions (from the logic model workshop)

The developers had the following expectations regarding the delivery of parental information sessions.

- Aim and format: it was planned for schools to engage parents through hosting an information session and sending a series of text messages. The developers provided guidance for running the information sessions and also a text messaging script which outlined suggested content.

- Beneficiaries: all Year 10 parents were invited to attend the parental information session, however, the developers acknowledged it was unlikely for previously disengaged parents to attend. The developers felt the text messages had the potential to reach a higher proportion of parents.

3.5.5.2. Evidence of implementation fidelity

The parental information session took place in most schools. Often, the parental information session for SSTT was combined with information about other subjects as those in the senior leadership team wanted to make the most of the time with parents in school. Heads of science felt this contributed to a high attendance from parents. During the parental information sessions, the head of science delivered the toolkit and talked through the slides from the developers. Whilst some schools taught parents about all five tools, other schools selected those they felt would be used the most by parents. For example, one of the schools did not include pre/post tests as it felt that this was not a strategy that could be used by parents/carers at home.

Schools that were not able to host the parental information evening instead informed parents about the toolkit during regular parents evenings. As the meetings typically only lasted for five minutes there was not enough time to go through each of the tools in detail. Therefore, parents/carers were also given hard copies of the toolkit resources. During the parents evening, one school also displayed the toolkit animations in the main hall for parents/carers to view.

3.5.5.3. Participants' experiences of the parental information sessions

Teachers felt that parents often struggle to support their children with their school work. Thus, they thought that parents would be receptive to joining this session to be able to understand the toolkit as well as learning how to use it at home with their children. In support, teachers perceived the sessions to be well attended by parents.

Teachers reported parents were positive about the information session. In particular, parents felt it was helpful to learn about strategies they could use at home to support their children without needing to understand the science content themselves.

Parents' views of the text messages were mixed. Some heads of science felt the use of text messages was useful to supplement the information taught at the parental information session.

> *'So if the kid had sat there saying, "I don't know what to do to revise", they could say, "well, remember the tools that we looked at when I went in for parents' evening: let's just watch that clip and think about the different tools and see how you could be using them" at home and they've got some support there to help their children' (head of science #1, post-intervention).*

In contrast, other heads of science felt the texting schedule was too intense, based on previous feedback from parents about the quantity of text messages sent to them by the school.

> *'I felt sending a text, one round of texts introducing the idea and then another round of texts saying why it might be good and then another round of texts of what to do, I think that would put parents off, but I think getting parents in would be good' (head of science #6, post-intervention).*

3.6. Costs

We estimated two types of cost: financial costs and time costs. The costs presented below relate to (1) hub school staff time and expenses for coordinating the delivery of SSTT at the hub and pilot level and (2) individual school staff time and school expenses for implementing SSTT.

The estimates are based on the numbers provided by the developers. There are two estimates of cost per student, one relates to implementing SSTT, the other to being a hub school and having to organise and deliver training. Both estimates include the cost of material resources necessary for implementation as well as the cost of school staff and supply cover. They do not include cost of prerequisites that do not represent a marginal cost for participating in the intervention.

3.6.1. Cost of coordinating the intervention at the hub and national level

Table 5 estimates the cost borne by a hub school for coordinating SSTT training and delivery to four other schools. Figures provided indicate that coordination and training delivery, on average, costs a hub school £27,712 or about £40 per participating student.

Table 5: Cost of coordination at the pilot and hub level for three hubs

Item	Cost to one hub school (£)[12]
Recruitment materials	300
Travel to hub schools	1,500
Venue hire (school-based)	1,750
Hospitality	1,750
Printing	200
Staffing—Strategic Development Lead	4,875
Staffing—Project Lead	10,237
Coaching	2,000
Staff—administrator	1,800
Head of science cover	1,800
Science teacher cover	1,500

[12] Figures may not add up due to rounding.

TOTAL per hub school	**27,712**
Average number of students per hub	**700**
Cost per student	**40**

In terms of school staff time, our estimates include time allocated to planning, travelling, and delivering activities for the training, workshop, and twilights. Overall, each hub school spent a total of 94 hours on planning, travelling, and delivering activities, including 33 hours of liaison.

Table 6: Hours spent coordinating the pilot

Activity	Number of hours per school hub[13]
Briefing twilight	8 (2)
Workshop 1	16 (6)
Coaching 1	12 (8)
Workshop 2	16 (6)
Mid-point review twilight	8 (2)
Parent workshop	14 (1)
Coaching 2	12 (8)
'Next step' twilight	8 (2)
Total number of hours	94 (33)

3.6.2. Staff and financial cost for individual schools

We also estimated staff time spent in engagement and support activities for individual schools participating in this programme. This includes attending the training, twilight sessions, and parent workshops mentioned above. In total, 30 hours was spent in engagement and support activities over a six-month period (see Table 7).

Table 7: Number of staff and time spent in engagement and support activities

Activity	Number of hours	Number of participating staff
Briefing twilight	3	2
Workshop 1	8	2
Coaching 1	1	1
Workshop 2	8	2
Mid-point review twilight	3	2
Parent workshop	2	2
Coaching 2	1	1
'Next step' twilight	4	2
Total number of hours	30	No total

In terms of staff time needed to deliver the intervention, our estimate was based on the assumption that the teachers would spend 20 minutes per tool per lesson during two lessons per week over 21 weeks (1.5 term). This estimate is crude as a teacher can refer to the tool throughout a lesson without actually using it. It was estimated that teachers spent 21 hours delivering the tool with students, 21 hours preparing it, and 7 hours checking student-made resources. Overall, 12 staff were involved in delivering the intervention per school.

[13] Numbers outside of brackets are hours spent planning, travelling, and delivering activities, whilst the numbers within brackets are the hours spent contacting school participants.

Table 8: Cost of implementing SSTT per student

Item	Cost to school (£)			Total cost over three years	Average cost per year (over three years)
	Year 1	Year 2	Year 3		
Head of science cover and travel cost	2,000	0	0	2,000	667
Science teacher cover and travel cost	1,700	0	0	1,700	567
Parent event	500	500	500	1,500	500
Materials (paper, card)	300	300	300	900	300
TOTAL	**4,500**	**800**	**800**	**6,100**	**2033**
Average number of students participating in a school[14]	**140**	**140**	**140**	**140**	**140**
Cost per student	**32**	**6**	**6**	**44**	**15**

The upfront cost of implementing the intervention in a school is £32 per student. The cost decreases to £6 per student in subsequent years. We assumed no teacher turnover, that is, the cost of training one head of science and a science teacher would only apply the first year.[15] We also assumed that parental information sessions are run annually to inform parents of the new Year 10 cohort and that new material is bought for each year for the new cohort. The estimates in Table 8 only include school costs, not those incurred by parents, students, or anyone else not employed by the school.

Following EEF guidance on cost analysis, we also estimated the average cost per year of the intervention if it is repeated over three years. This means that the estimated cost per student for implementing SSTT over three years is £44, that is, £15 per student per year over three years.

3.6.3. Overall assessment

Overall, the cost per student for a hub school involved in coordinating and training is £40 whilst for a non-hub school implementing SSTT it is £15. Both these costs can be considered as low. According to the EEF Toolkit, most meta-cognition and self-regulation projects funded by the EEF were estimated as costing under £80 per student, including the necessary professional development for teachers.[16]

3.7. Readiness for trial

We assessed the SSTT's readiness for trial based on the following criteria: credibility of the logic model, support for the intervention, cost, innovativeness of the intervention, timing and duration of the trial, dosage, capacity of the developers to deliver, and responsiveness to the evaluation.

Our overall assessment is that, despite the strengths of the intervention, further consideration should be given to a number of key issues before moving on to a trial. The following section summarises our main arguments.

3.7.1. Arguments in favour of proceeding to a trial

3.7.1.1. Logic model

[14] The average number of pupils per school is calculated from the total number of participating pupils in the pilot (1,200) divided by the number of schools participating in the pilot (15).
[15] Teachers in England complete an average of six years before either moving school, leaving teaching, or retiring (Allen, R., Burgess, S. and Mayo, J. (2012) 'The Teacher Labour Market, Teacher Turnover and Disadvantaged Schools: New Evidence for England', CMPO Working paper, 12/294).
[16] https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit/meta-cognition-and-self-regulation/

We found that the logic model underpinning the intervention was sound. We used the pre-intervention logic model workshop to get clarifications on some of the links in the logic model and to test some of the developers' assumptions. We decided not to modify the logic model further to this meeting.

During the post-intervention workshop, we discussed some of the emerging findings of the evaluation and reviewed the logic model accordingly. The developers indicated that, in a trial, very minor modifications would need to be made as follows.

- Project timeline: the developers felt extending the intervention across one full school year would be beneficial for delivery. In particular, they felt this revised timeline would give teachers more time to familiarise themselves with the tools and embed them in their teaching practice.
- Additional support activities: in a longer trial, recruitment would need to start in the spring and delivery in September. This longer recruitment period would justify the organisation of a pre-launch event in June aimed at making it very clear what the project involves and what the school needs to do in order to take part. This event would need to be attended by a member of the school leadership team who may not necessarily be a science teacher; SLT support is important for successful implementation.
- Use of the logic model: the developers also felt a timeline and logic model provided to teachers would aid implementation as they would understand what needs to be done when in order to successfully embed the tools.

The revised logic model can be found in Appendix II.

3.7.1.2. Support for the intervention

Overall, we found that support for the intervention was high, not only among heads of science and teachers (see sections 3.3.2) but also among a majority of students who found that the SSTT had had a positive effect on memorisation (see section 3.3.3). In a trial, such a high level of engagement would facilitate recruitment and implementation fidelity/compliance.

3.7.1.3. Cost

Given the low unit cost of the intervention, our assessment is that a trial would be scalable (see section 3.6).

3.7.2. Issues requiring further consideration

3.7.2.1. Usual practice

In a trial, the developers would need to clearly demonstrate the innovativeness of the intervention. Many of the teachers we interviewed had used one or more of the strategies in the toolkit—although not as *self-testing strategies*. For example, we found that mind maps are frequently used in some schools but not *mind-mapping tests*. We also found that two thirds of students have used flash cards in the past and one third have used mind maps. Thus, in a significant number of schools, an impact evaluation would compare the effect of 'mind-mapping tests' vs 'standard mind maps'. This is likely to dilute the effect of the intervention.

3.7.2.2. Timing and duration of the trial

The ideal target year group of the SSTT and length of intervention delivery remain to be discussed. In an efficacy trial, the outcome of interest would most likely be the grade achieved in science GCSE at the end of Year 11. Yet the SSTT was piloted in Year 10. The developers picked Year 10 for the pilot because it guaranteed all students would be studying for GCSEs. However, developers agreed with the possible benefits of starting the intervention in Year 9 (for example, better familiarity with the tools). Several teachers indicated that the SSTT should be introduced in Year 9. However, this would make the trial very long.

3.7.2.3. Dosage

In a trial, the developers would need to set out clear and realistic expectations in terms of dosage (that is, the frequency and duration of use of the SSTT). In the pilot, we found out that dosage (1) varied significantly between schools, which indicates different degrees of compliance, and (2) was reported to be higher among teachers than among students. This discrepancy should be interpreted with caution as we used different data-collection methods with these two groups (student survey vs teacher interviews) and did not attempt within-school comparisons. These two findings suggest that participating teachers and students were not very clear about the expected dosage or did not find the expected dosage optimal.

3.7.2.4. Developers' capacity

It remains unclear whether the developers would have the capacity to implement a trial. We have identified three challenges. The first challenge is the recruitment of a large number of settings. During the post-intervention logic model workshop, the developers expressed doubts about their capacity to recruit schools beyond their existing network and in a relatively tight timeframe. The second challenge is the scaling up of support activities to schools (workshops, coaching, twilight sessions etc.). Whilst we noted a high level of implementation fidelity in the pilot, whether the developers could step up their efforts in the context of a trial is not assured, not least because one of the three project leads has taken up a new role since the termination of the project. The third challenge is the capacity of the developers to work closely with the evaluators and to provide information in a timely manner.

3.7.2.5. Teachers' responsiveness to the evaluation

A trial would require a high level of engagement with the evaluation among teachers and heads of science. This was not achieved in the pilot.

- Several heads of science did not seem to be aware—or to recall—that the intervention would be evaluated when we sent them a Memorandum of Understanding.
- One school dropped out of the evaluation at the start of the project because of the perceived burden of research activities on school staff.
- We had to adjust our teacher sampling approach given a lack of engagement from some heads of science and teachers.
- One school cancelled its school visit. The reason given was that it was too close to the end of term and it would not have time to fit it into its programme.
- We had to cancel all pre-intervention teacher interviews and a significant number of observations.
- Three schools out of 12 did not complete the student survey. This suggests that the heads of science in these schools did not forward the link to students or did not encourage them to take part.
- Generally, it was difficult to get timely responses to our emails and phone calls.

Conclusion

Key findings

Research question	Finding
Is there evidence to support the theory of change?	Students responded well to the intervention, but some strategies were perceived as more effective and enjoyable than others. Students with higher self-efficacy had a more positive experience than students with lower science self-efficacy. Some strategies in the SSTT are close to usual practice (flash card revision).
Is the approach feasible to deliver?	Heads of science and science teachers responded well to the intervention, both in terms of attitudes and behaviours. Support activities were faithfully implemented and the cost of SSTT is low.
Is the intervention ready to be evaluated in a trial?	Despite the above-mentioned strengths, further consideration must be given to a few issues to make a trial feasible as the conditions of implementation were more favourable than they would be in a trial. These issues include (1) the innovativeness of the intervention, (2) clarity over the target year group and optimal dosage, (3) the optimal dosage of the intervention, (4) the developers' willingness and capacity to recruit and deliver on a greater scale, and (5) the responsiveness of schools and teachers to the evaluation.

Despite a sound logic model, a high level of implementation fidelity from the developers, a positive response from both teachers and students, and a low cost, our conclusion is that a number of issues require further consideration before a trial can be deemed feasible.

Firstly, the innovativeness of the SSTT must be emphasised. Many teachers and students told us that they had already used some of the SSTT tools before the pilot. While teachers recognised the benefits of the 'self-testing' approach and of combining these tools into a single kit, it was not clear how innovative they found the SSTT. The insufficient contrast between intervention and control reduces the likelihood of detecting an impact.

Secondly, the optimal timing and duration of a trial deserve further consideration. There was widespread support among teachers for an introduction of the SSTT in Year 9 rather than Year 10, as in the pilot. There is evidence that some teachers used the SSTT in Year 9 during the pilot. However, in a trial, the most likely outcome would be the grade achieved in science GCSE at the end of Year 11. This would make the intervention delivery, the trial—or both—very long. The benefits of a longer trial would have to be assessed against (1) the cost of such a trial and (2) the greater risk of attrition.

Thirdly, the optimal dosage must be clarified and communicated to teachers. This will require managing the tension between effectiveness and feasibility. A high dosage is likely to have a larger effect on memorisation and attainment but might be hard to implement consistently. Conversely, a low dosage might be easier to implement but may not prove effective. The evidence collected from the student survey and the teacher interviews suggests that the expected dosage was not clear or considered optimal—or both.

Fourthly, the developers must demonstrate their willingness and capacity to (1) recruit a large number of schools, (2) scale up delivery, and (3) meet the evaluators' requirements for a trial.

Lastly, the developers and evaluators need to find a way of making participating schools and teachers more responsive to the evaluation. This was a key issue in the pilot, possibly because teachers considered the SSTT very similar to their usual practice. In a trial, it would be crucial for school staff to remain engaged in the evaluation, which may prove to be challenging.

Limitations of the evaluation

It is safe to assume that the views expressed by heads of science and teachers in this report are somewhat more favourable to the SSTT than those of the 'average' teacher. There are two main reasons for this. First, the revised sampling methodology of the evaluation resulted in an oversampling of heads of science (who volunteered to take part in the pilot) and self-selected teachers. Second, the high number of cancelled/unproductive teacher interviews is also likely to have skewed the results.

Two schools also did not take part in the student survey, perhaps a result of being less engaged with the pilot.

Future research

This evaluation did not address all the research questions included in the developers' logic model. Future research is needed to explore:

- students' understanding of the benefits of self-testing;
- parents' understanding of the benefits of self-testing; and
- the perceived effect of the SSTT on parents' confidence in their ability to support their child.

References

Adesope, O. O., Trevisan, D. A. and Sundararajan, N. (2017) ‘Rethinking the Use of Tests: A Meta-Analysis of Practice Testing’, *Review of Educational Research*, 87 (3), pp. 659–701. DOI: 10.3102/0034654316689306

Currie, C. et al. (eds) (2012) ‘Social Determinants of Health and Well-Being Among Young People: Health Behaviour in School-Aged Children (HBSC) Study: International Report from the 2009/2010 Survey’, World Health Organisation, Health Policy for Children and Adolescents No. 6. http://www.euro.who.int/__data/assets/pdf_file/0003/163857/Social-determinants-of-health-and-well-being-among-young-people.pdf

Dunlosky, J. et al. (2013) ‘Improving Students’ Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology’, ***Psychological Science in the Public Interest***, 14 (1), pp. 4–58. DOI: 10.1177/1529100612453266

Education Endowment Foundation (2015) ‘EEF Guidance on Cost Evaluation’, April 2015.

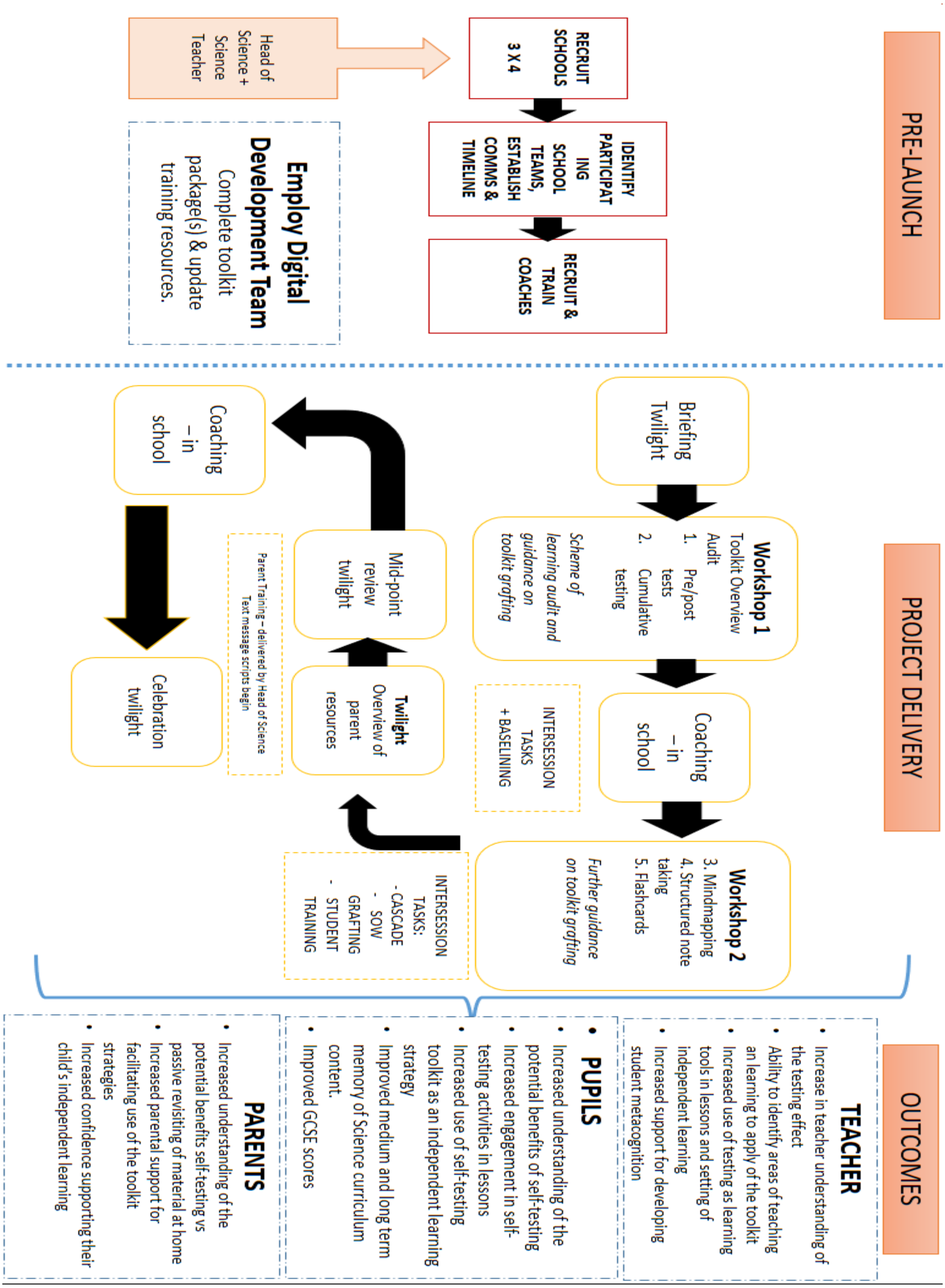
McCaleb-Kahan, P., Wenner, R. (2009) ‘The Relationship of Student Demographics to 10th Grade MCAS Test Anxiety’. Paper presented at the 40th Annual Meeting of the Northeastern Educational Research Association, Rocky Hill, CT, October 2009. https://pdfs.semanticscholar.org/5af3/f6e2105defe784c383753772a67dc5f8067c.pdf

Pan, S. C. and Rickard, T. C. (2018) ‘Transfer of Test-Enhanced Learning: Meta-Analytic Review and Synthesis’, *Psychological Bulletin*, 144 (7), pp. 710–756. DOI: 10.1037/bul0000151

Quigley, A., Muijs, D. and Stringer, E. (2018) ‘Metacognition and Self-Regulated Learning: Guidance Report’. https://educationendowmentfoundation.org.uk/tools/guidance-reports/metacognition-and-self-regulated-learning

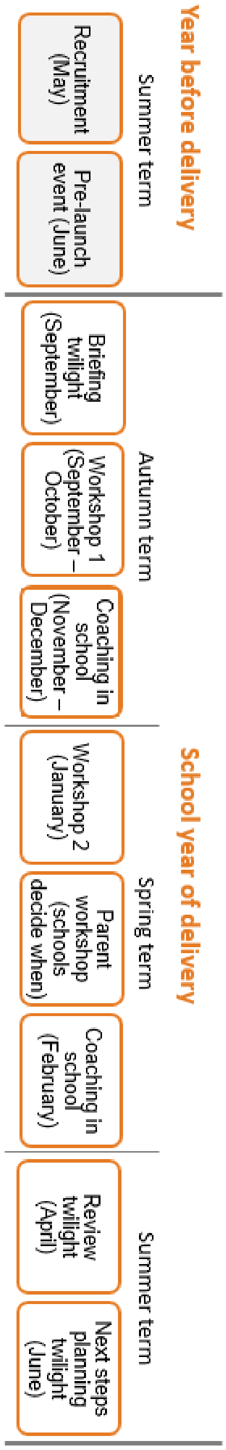
Appendix I. Original logic model

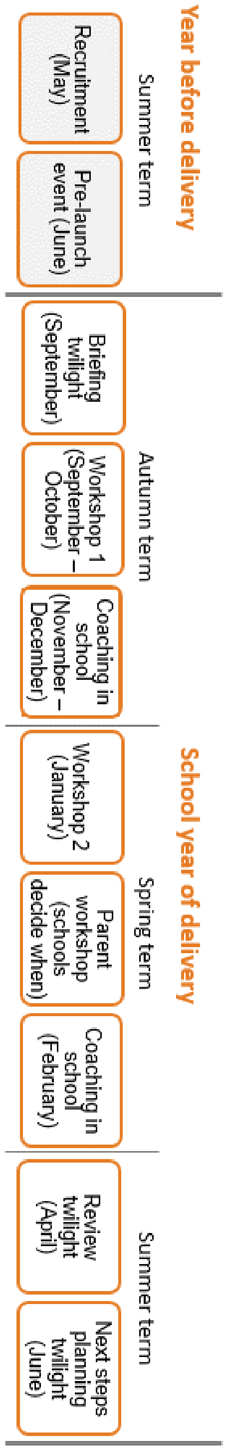
The original logic model was created by the developers. During the pre-intervention workshop, it was concluded the logic model was sound and did not require any modifications.



Appendix II. Revised logic model

During the post-intervention logic model workshop, the developers concluded that the pre-launch and outcomes in the original logic model should remain unchanged. However, the developers felt the project delivery should be changed to reflect an extended delivery timeline over one academic year.





Appendix III. Survey questionnaire

Science Self-Testing Toolkit Student Survey

Questions are documented as follows:

{Question routing- who is asked the question}

Question Name

Question text

: Question response options

(Variable label)

i. Introduction (n.b all titles for reference, not scripted)

Thank you for agreeing to take part in this survey!

This is a short survey about how you revise for science tests and exams. Over the past few months you may have been introduced to the 'science self-testing toolkit' ("the toolkit") in your science classes. If you have, we will also ask some questions about this.

Please answer about all of your science lessons including biology, chemistry, and physics. It doesn't matter whether you are in Double or Triple science.

This survey will take about 10 minutes to complete. Please answer as honestly as you can. This is not a test or judgement of your study habits; there are no right or wrong answers!

Your answers will be anonymous, kept private and will not be shown to anyone at your school. Only the research team will see your responses and we will store them confidentially. Thank you very much for your help.

If you want to be entered in the draw for a high street shopping voucher, make sure to enter your email at the end of the survey. We will be drawing for 6 prizes:

1 x £50.00 voucher

2 x £25.00 voucher

3 x £15.00 voucher

The vouchers can be used at over 150 stores and restaurants including Schuh, Topshop/Topman and Pizza Express! Please tick below and press 'Next' to continue

: Tick box: I am happy to continue with the survey

ii. Background

{ASK ALL}

1. School

What is the name of your school?
[drop down. Single select]

[redacted]

Variable label: School

{ASK ALL}

2. Sex

Are you male or female?

1. Male
2. Female
3. Other
4. Prefer not to say

Variable label: sex

iii. Attitudes to learning and memorisation

Intro text (may need to script as question text at effort)

Next are some things that students have said about learning and studying GCSE science. For each statement please say how strongly you agree or disagree with it. Please think about all of your science classes as a whole when answering.

{ASK ALL}

3. Effort

I work as hard as I can in science

1. Strongly agree
2. Agree
3. Neither agree or disagree
4. Disagree
5. Strongly disagree

Variable label: effort at school

{ASK ALL}

4. Marks

I generally get good marks for my work on science tests and exams.

1. Strongly agree
2. Agree
3. Neither agree or disagree
4. Disagree
5. Strongly disagree

Variable label: self-reported attainment

{ASK ALL}

5. Enjoy

I enjoy science lessons.

1. Strongly agree
2. Agree
3. Neither agree or disagree
4. Disagree
5. Strongly disagree

Variable label: whether enjoys science

{ASK ALL}

6. MemoriseE

I find it easy to memorise facts and concepts for science tests and exams.

1. Strongly agree
2. Agree
3. Neither agree or disagree
4. Disagree
5. Strongly disagree

Variable label: whether memorisation easy

{ASK ALL}

7. MemoriseS

I feel relaxed and confident when memorising facts and concepts for science tests and exams.

1. Strongly agree
2. Agree
3. Neither agree or disagree
4. Disagree
5. Strongly disagree

Variable label: whether memorisation stressful

iv. Prior to intervention

Intro text (may need to be scripted as question text at revision)

In the next few questions, we would like to know how you revised for science in the **Year 9**. It may be helpful to think back to a science test you took in Year 9 when answering the next few questions.

These questions will not be used to judge your revision habits. We just want to know if and how your revision habits have changed over time.

{ASK ALL}

8. PStrategy

Which memorisation and revision strategies did you use most often to memorise and revise for your science tests or exams in Year 9?

Choose **up to three** strategies you regularly used in Year 9.

{multicode}

- Making mind maps using class notes, or from memory
- Creating posters or diagrams
- Thinking of questions for each topic and regularly testing yourself on all the topics you've covered
- Testing what you know about a topic before you cover it in class, and again once you've learnt it
- Doing practice tests or past papers
- Answering quiz questions with friends or family
- Re-doing homework questions
- Copying out study notes
- Making flash cards from notes
- Using flash cards to test yourself
- Highlighting class notes
- Writing down questions as you learn, and then trying to answer these without looking at your notes
- Re-reading class notes and textbooks (but not highlighting or copying)
- Speaking notes out loud
- Revision websites (e.g. BBC Bitesize)
- Watching video clips
- Quizzing apps (e.g. EducationQuizzes)
- Other (Please explain)
- None of these (EXCLUSIVE)

{ASK IF revision =Yes}

10. PEffectiveS

Thinking now about your Year 9 revision/learning strategies overall – how helpful do you think your approach was for **remembering** a science lesson for **a test or exam on that topic**?

- Very unhelpful
- Unhelpful
- Not helpful or unhelpful
- Helpful
- Very helpful

Variable label: Effectiveness of prior strategies: short term

{ASK IF revision =Yes}

11. PEffectiveLT

Thinking now about your Year 9 revision/learning strategies overall – how helpful do you think your approach was for **remembering** a science lesson **and recalling that information over a longer period of time?** - For example, after a test?

- Very unhelpful
- Unhelpful
- Not helpful or unhelpful
- Helpful
- Very helpful

Variable label: effectiveness of prior strategies: long term

{ASK IF revision =Yes}

12. PUnderstanding

Thinking now about your Autumn term revision/learning strategies overall – how helpful do you think your approach was **understanding** the ideas and concepts of a science lesson?

- Very unhelpful
- Unhelpful
- Not helpful or unhelpful
- Helpful
- Very helpful

Variable label: effectiveness of prior strategies: understanding

{ASK IF revision =Yes}

PTime

Overall, how time consuming do you think your approach was?

- Very time consuming – It took me a very long time to get through a small amount of material
- Quite time consuming, It took quite a long time studying to get through not much material
- Not very time consuming, I could get through quite a lot of material in an OK amount of time
- Not time consuming at all, I could get through a lot of material very quickly

Variable label: how time consuming previous strategies were

{ASK IF revision =Yes}

13. Continuous

Some students revise material that they have learned immediately after they have learned it and may continue to revise that material up to a test or exam on the subject. Others revise all in one go before the test or exam.

What is most similar to what you normally do in science?

- I continuously revise lessons that I have learned
- I revisit older topics once in a while
- After I learn a lesson I normally do not revise the material until I am studying for a test or exam.

Variable label: whether revise continuously

v. The intervention

In the Spring term you may have been taught some new memorisation and revision strategies in your science classes. These come from the Science Self-Testing Toolkit (SSTT).

This next section asks about whether you have used the strategies and what you thought about them.

If none of the strategies sound familiar, that is okay! Depending on your science teacher and school, you may not have used any yet. Below is a description of each strategy in the toolkit. We will then ask you if you have been introduced to any of these.

{ASK ALL}

The strategies in the toolkit are:

1. Pre-tests/Post-tests. Doing a test before you have learned the lesson or topic. Once you have learned the lesson or topic, you do the test again.

2. Cumulative quizzing. Coming up with questions about a topic you are learning and adding to the list of questions over time, testing yourself on each question each time. Your teacher may give you the questions, you may make them up as a class, or you make them up on your own.

3. Flashcards. Making flashcards (either on paper, the computer, or an app). Over time, you shuffle the order of the deck of cards and remove some cards for a few days before adding them back in. Your teacher may give you the flashcards, you may make them as a class, or you may make them up on your own.

4. Structured note-taking. Writing down questions as you write or read notes in the margin of your page and then trying to answer the questions without looking at your notes. Your teacher may give you the questions, you may make them up as a class, or you make them up on your own.

5. Blind mind mapping. Creating a mind map for a topic you have learned, tracing the structure of the map on a blank piece of paper, and attempting to fill in the information whilst looking at your notes as little as possible. Your teacher may give you structured mind map, you may make it up as a class, or you make it up on your own.

{ASK ALL}

14. Used

Which, if any, of these 5 toolkit strategies have you been introduced to in your spring term science classes?

[multicode]

1. Pre-tests/post-tests
2. Cumulative quizzing
3. Flashcards
4. Structured note taking
5. Blind mind mapping
6. I haven’t been introduced to any of these (Exclusive)

Variable label: whether used SSTT

{ASK IF Used <> 6}

15. HowMuch

Roughly how many times have you used any one of these strategies in class?

- Every class
- Every other class
- Once every few classes
- In just one or two classes
- I have only used one of the strategies once in my science classes

Variable label: how often used SSTT

{ASK IF USED <>6}

16. STTHome

Have you used any of the toolkit strategies outside of lessons? [MULTICODE]

Choose all that you have tried.

1. Pre-tests/Post-tests.
2. Cumulative quizzing.
3. Flashcards.
4. Structured note-taking.
5. Blind mind mapping.
6. I have not tried any of the strategies outside of lessons (EXCLUSIVE).

Variable label: how often used SSTT

{ASK IF STTHome = 6}

17. WNhome

Why have you not used any one of these toolkit strategies outside of lessons? [multicode]

Choose all that apply.

I already have memorisation and revision methods I like more
I am unsure of how to use these strategies at home
I do not think these strategies will help me
It is too much effort to use these strategies at home
I do not have enough time
I did not know I could use the strategies at home
I was not told to use the strategies at home
I haven't done any revision/studying at home since learning the strategies
Other (please explain)

Variable label: why haven't used SSTT outside of classes

{ASK IF STTHome <>6}

18. FrequencyH

Roughly how often have you used any of these toolkit strategies outside of lessons?

Every day
3-5 times a week
1-2 times a week
3-4 times a month
1-2 times a month

Less than once a month
I have only used the strategies once at home

Variable label: how often used at home

{ASK IF Used <> 6}

19. STTShort

Thinking about the toolkit strategies you have used at home or at school, how helpful or unhelpful was it for **remembering** a science lesson for **a test or exam on that topic**?

Very unhelpful
Unhelpful
Not helpful or unhelpful
Helpful
Very helpful

Variable label: effectiveness of SSTT short term

{ASK IF Used <> 6}

20. STTlongT

Thinking about the toolkit strategies you have used at home or at school, how helpful or unhelpful was it for **remembering** a science lesson over a longer period of time? For example, still remembering it after a test.

Very unhelpful
Unhelpful
Not helpful or unhelpful
Helpful
Very helpful

Variable label: effectiveness of SSTT long term

{ASK IF Used <> 6}

21. STTunder

Thinking about the toolkit strategies you have used at home or at school, how helpful or unhelpful was it for **understanding** the ideas and concepts of a science lesson?

Very unhelpful
Unhelpful
Not helpful or unhelpful
Helpful
Very helpful

Variable label: effectiveness of SSTT - understanding

{ASK IF Used <> 6}

22. StratEnjoy

Of the toolkit strategies that you have used, which one do you **enjoy** using the most?

Pre-tests/Post-tests.
Cumulative quizzing.
Flashcards.

Structured note-taking.
Blind mind mapping.
I do not enjoy any strategy I have used.

Variable label: most enjoyable strategy

{ASK IF Used <> 6}

23. StratRem

Of the toolkit strategies that you have used, which one do you think is the most helpful for **remembering** a science lesson?

Pre-tests/Post-tests.
Cumulative quizzing.
Flashcards.
Structured note-taking.
Blind mind mapping.
I do not think any strategy is helpful for remembering a lesson.

Variable label: most helpful strategy for remembering

{ASK IF Used <> 6}

24. StratUn

Of the strategies that you have used, which one do you think is the most helpful for **understanding** a science lesson?

Pre-tests/Post-tests.
Cumulative quizzing.
Flashcards.
Structured note-taking.
Blind mind mapping.
I do not think any strategy is helpful for understanding a lesson.

Variable label: most helpful strategy for understanding

{ASK IF Used <> 6}

25. STStress

How stressful do you find using the toolkit strategies, or do you not find them stressful at all?

Very stressful
Fairly stressful
A little bit stressful
Not stressful at all

Variable label: whether found SSTT stressful

{ASK IF used <> 6}

26. FutureS

Would you like to continue using the toolkit in your science classes next school year?

Yes
No
I don't know

Variable label: whether want to continue using in class

{ASK IF used <> 6}

27. Others

Do you think the toolkit strategies would be helpful to understand and revise other subjects (e.g. maths or English)?

Yes
No
I don't know

Variable label: whether SSTT would be helpful for other subjects

{ASK ALL}

28. Lottery

If you would like to be entered in the draw to win a high street voucher worth up to £50, please enter your email.

:text fill

End

That's the end of the survey. Thank you very much for taking the time to answer these questions.

Appendix IV. Descriptive statistics relating to the use of SSTT

Types	Categories		Proportion (%)
Previous strategies	Use	Mind map from notes	39.00
		Pretest/posttest	6.00
		Flash cards from notes	37.50
		Flash cards to test self	27.00
		Structured note taking	5.00
	Effectiveness	Helpful or very helpful for understanding ideas and concepts of science lessons	47.30
		Helpful or very helpful for short term memory	35.00
SSTT	Use	Flashcards	76.90
		Blind mind mapping	42.50
		Pretest/posttest	36.00
		Structured note-taking	29.50
		Cumulative quizzing	29.30
		Not used any	3.50
		Not sure/prefer not to say	6.30
	Frequency of use in class	Every class	10.30
		Every other class	17.40
		Once every few classes	33.00
		In just one or two classes	22.40
		Once	7.90
	Ever used at home	Flashcards	72.80
		Structured note-taking	26.00
		Blind mind mapping	25.90
		Pretest/posttest	19.80
		Cumulative quizzing	14.60
		Not used any	11.50
		Not sure/prefer not to say	2.70
	Frequency of use outside of class	Every day	2.20
		35 times a week	10.40
		12 times a week	25.80
		34 times a month	15.90
		12 times a month	19.30
		Less than once a month	13.20
		Once	6.00
	Helpfulness	Helpful or very helpful for short term memory	58.30
		Helpful or very helpful for long term memory	50.40
		Helpful or very helpful for understanding ideas and concepts of science lessons	54.40
	Stressfulness	Very or fairly stressful	30.30
	Most enjoyable	Flashcards	50.30
		Structured note-taking	11.50

		Blind mind mapping	11.50
		Pretest/posttest	5.60
		Cumulative quizzing	5.60
		Do not enjoy any	10.20
Most helpful	For remembering	Flashcards	46.20
		Structured note-taking	14.60
		Blind mind mapping	11.50
		Cumulative quizzing	8.10
		Pretest/posttest	7.60
		Do not enjoy any	5.40
	For understanding lessons	Flashcards	36.70
		Structured note-taking	23.80
		Blind mind mapping	11.10
		Pretest/posttest	10.40
		Cumulative quizzing	7.30
		Do not think any strategy is helpful	6.80
Future use	Want to continue in class	Yes	61.80
	Helpful to revise other subject	Yes	64.10

Appendix V. Association between science self-efficacy and experience relating to SSTT

Types	Categories		Self-efficacy		p-value
			Average and above (%)	Below average (%)	
Effectiveness of prior strategies	Short-term	Very unhelpful	6.64	9.93	0.000
		Unhelpful	8.92	17.02	
		Not helpful or unhelpful	19.29	34.75	
		Helpful	51.45	25.30	
		Very helpful	11.00	2.84	
		Don't know/prefer not to say	2.70	10.17	
	Long-term	Very unhelpful	4.60	12.11	0.000
		Unhelpful	13.39	26.37	
		Not helpful or unhelpful	28.87	33.73	
		Helpful	43.31	17.81	
		Very helpful	6.90	1.43	
		Don't know/prefer not to say	2.93	8.55	
	Understanding	Very unhelpful	4.20	6.90	0.000
		Unhelpful	6.93	20.95	
		Not helpful or unhelpful	23.11	31.43	
		Helpful	53.15	29.29	
		Very helpful	10.50	2.38	
		Don't know/prefer not to say	2.10	9.05	
Effectiveness of SSTT	Short-term	Very unhelpful	2.75	6.96	0.000
		Unhelpful	5.50	10.86	
		Not helpful or unhelpful	19.04	33.15	
		Helpful	54.59	38.16	
		Very helpful	14.91	4.18	
		Don't know/prefer not to say	3.21	6.69	
	Long-term	Very unhelpful	3.95	7.02	0.000
		Unhelpful	5.81	15.17	
		Not helpful or unhelpful	21.63	35.67	
		Helpful	53.02	31.74	
		Very helpful	10.93	3.93	
		Don't know/prefer not to say	4.65	6.46	
	Understanding	Very unhelpful	2.80	4.79	0.000
		Unhelpful	6.76	12.68	
		Not helpful or unhelpful	18.88	36.34	
		Helpful	57.11	34.37	
		Very helpful	10.26	5.63	
		Don't know/prefer not to say	4.20	6.20	

Types	Categories		Self-efficacy		p-value
			Average and above (%)	Below average (%)	
SSTT usage	Whether revise continuously	I continuously revise	10.23	5.52	0.000
		I revisit older topic	44.35	31.18	
		After I learn a lesson	42.00	49.16	
		Don't know/prefer not to say	3.41	14.15	
	Frequency of SSTT usage	Every day	3.44	0.65	0.000
		3-5 times a week	12.29	8.17	
		1-2 times a week	30.47	20.26	
		3-4 times a month	18.67	11.44	
		1-2 times a month	16.71	23.20	
		Less than once a month	10.32	17.65	
		I have only used the tool once	4.18	10.13	
		Don't know/prefer not to say	3.93	8.50	
	Most enjoyable strategy	Pre-tests/Post-tests	5.62	5.07	0.328
		Cumulative quizzing	5.85	4.79	
		Flashcards	53.40	50.14	
		Structured note-taking	11.48	10.99	
		Blind mind mapping	11.94	10.99	
		I do not enjoy any strategy	7.49	12.68	
		Don't know/prefer not to say	4.22	5.35	
Most helpful strategy	For remembering	Pre-tests/Post-tests	7.78	6.84	0.032
		Cumulative quizzing	8.25	7.98	
		Flashcards	46.70	46.15	
		Structured note-taking	16.27	12.82	
		Blind mind mapping	12.03	10.83	
		I do not enjoy any strategy	2.83	8.55	
		Don't know/prefer not to say	6.13	6.84	
	For understanding	Pre-tests/Post-tests	9.52	10.86	0.191
		Cumulative quizzing	8.33	5.71	
		Flashcards	34.29	34.00	
		Structured note-taking	26.43	21.71	
		Blind mind mapping	10.95	11.71	
		I do not enjoy any strategy	5.71	8.00	
		Don't know/prefer not to say	4.76	8.00	
Enjoyment	How stressful	Very stressful	5.48	14.41	0.000
		Fairly stressful	17.38	23.63	
		A little bit stressful	45.00	38.33	
		Not stressful at all	25.95	14.99	
		Don't know/prefer not to say	6.19	8.65	
	Want to continue in class	Yes	74.22	50.87	0.000
		No	11.69	23.26	
		Don't know/prefer not to say	14.08	25.87	
		Yes	74.16	54.07	0.000

Types	Categories		Self-efficacy		p-value
			Average and above (%)	Below average (%)	
	Helpful to revise other subject	No	14.11	23.84	
		Don't know/prefer not to say	11.72	22.09	

Appendix VI. Association between gender and experience relating to SSTT

Types	Categories		Male (%)	Female (%)	p-value
Effectiveness of prior strategies	Short-term	Very unhelpful	8.75	7.35	0.073
		Unhelpful	9.54	15.64	
		Not helpful or unhelpful	26.04	26.07	
		Helpful	40.76	39.57	
		Very helpful	7.55	6.40	
		Don't know/prefer not to say	7.36	4.98	
	Long-term	Very unhelpful	8.43	5.95	0.001
		Unhelpful	14.86	25.48	
		Not helpful or unhelpful	31.12	31.67	
		Helpful	33.33	29.52	
		Very helpful	5.02	3.10	
		Don't know/prefer not to say	7.23	4.29	
	Understanding	Very unhelpful	5.02	4.33	0.095
		Unhelpful	11.65	17.07	
		Not helpful or unhelpful	26.71	28.61	
		Helpful	42.77	39.90	
		Very helpful	6.83	6.01	
		Don't know/prefer not to say	7.03	4.09	
Effectiveness of SSTT	Short-term	Very unhelpful	5.03	3.79	0.076
		Unhelpful	8.70	7.59	
		Not helpful or unhelpful	22.88	29.27	
		Helpful	46.22	47.97	
		Very helpful	10.98	8.40	
		Don't know/prefer not to say	6.18	2.98	
	Long-term	Very unhelpful	7.62	2.76	0.001
		Unhelpful	9.70	11.05	
		Not helpful or unhelpful	24.48	31.77	
		Helpful	42.73	43.92	
		Very helpful	7.62	7.18	
		Don't know/prefer not to say	7.85	3.31	
	Understanding	Very unhelpful	5.09	1.94	0.018
		Unhelpful	8.10	12.47	
		Not helpful or unhelpful	25.00	28.53	
		Helpful	46.30	46.26	
		Very helpful	9.26	7.20	
		Don't know/prefer not to say	6.25	3.60	
SSTT usage	Whether revise continuously	I continuously revise	7.76	6.75	0.002
		I revisit older topic	42.86	32.29	
		After I learn a lesson	39.80	52.53	
		Don't know/prefer not to say	9.59	8.43	

Types	Categories		Male (%)	Female (%)	p-value
	Frequency of SSTT usage	Every day	2.67	0.87	0.672
		3-5 times a week	11.20	10.17	
		1-2 times a week	25.07	26.74	
		3-4 times a month	17.07	15.12	
		1-2 times a month	18.67	20.64	
		Less than once a month	13.07	12.79	
		I have only used the tool once	6.67	6.69	
		Don't know/prefer not to say	5.60	6.98	
	Most enjoyable strategy	Pre-tests/Post-tests	8.58	2.50	0.000
		Cumulative quizzing	5.57	5.56	
		Flashcards	45.94	56.94	
		Structured note-taking	9.05	14.72	
		Blind mind mapping	11.37	9.72	
		I do not enjoy any strategy	12.99	6.67	
		Don't know/prefer not to say	6.50	3.89	
Most helpful strategy	For remembering	Pre-tests/Post-tests	8.86	6.46	0.195
		Cumulative quizzing	6.99	9.55	
		Flashcards	46.85	47.19	
		Structured note-taking	12.82	16.29	
		Blind mind mapping	10.49	11.24	
		I do not enjoy any strategy	6.29	3.65	
		Don't know/prefer not to say	7.69	5.62	
	For understanding	Pre-tests/Post-tests	10.35	11.05	0.127
		Cumulative quizzing	8.24	6.23	
		Flashcards	34.12	32.86	
		Structured note-taking	20.47	27.20	
		Blind mind mapping	12.71	9.92	
		I do not enjoy any strategy	5.88	7.65	
		Don't know/prefer not to say	8.24	5.10	
Enjoyment	How stressful	Very stressful	9.41	10.29	0.004
		Fairly stressful	17.18	23.71	
		A little bit stressful	40.00	44.57	
		Not stressful at all	24.71	16.57	
		Don't know/prefer not to say	8.71	4.86	
	Want to continue in class	Yes	59.95	66.00	0.057
		No	20.14	13.71	
		Don't know/prefer not to say	19.91	20.29	
	Helpful to revise other subject	Yes	63.42	67.43	0.357
		No	19.71	16.00	
		Don't know/prefer not to say	16.86	16.57	

Education Endowment Foundation

The Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
London
SW1P 4QP

https://educationendowmentfoundation.org.uk

@EducEndowFoundn

Facebook.com/EducEndowFoundn