Science Self-Testing Toolkit Pilot Evaluation Plan

NatCen Social Research
Arnaud Vaganay

PROJECT TITLE	Science Self-Testing Toolkit
DEVELOPER (INSTITUTION)	Kingsbridge Academy
EVALUATOR (INSTITUTION)	NatCen Social Research
PRINCIPAL INVESTIGATOR(S)	Arnaud Vaganay
EVALUATION PLAN AUTHOR(S)	Arnaud Vaganay, Sarah Frankenburg
PUPIL AGE RANGE AND KEY STAGE	Year 10 (Key Stage 4)
NUMBER OF SCHOOLS/ SETTINGS	12 schools
NUMBER OF PUPILS	2100

Evaluation plan version history

VERSION	DATE	REASON FOR REVISION
1.0	21 Feb 2019	

1. Intervention

1.1. Overview

The Science Self-Testing Toolkit (SSTT) is a suite of five activities that all aim to increase the amount of pupil self-testing in Key Stage 4 science study.

The pilot will run in 12 schools from January to July 2019.

1.2. Why

In recent years GCSE science exams have changed from termly modular exams throughout Y10 and Y11, to end of year exams in Y10 and Y11 to the current model, which is terminal exams at the end of Y11. This has significant implications for teachers and students:

- Students are required to memorise a huge amount of content over the two-year course;
- Science teachers need to adapt their teaching to support this need for improved memory retention and recall.

The SSTT addresses this challenge through self-testing (ST). ST is a teaching and learning technique that encourages pupils to engage in active retrieval of memories during revisions rather than relying on more passive approaches such as rereading material.

The effect of ST on learning outcomes is well documented. Overall, this effect seems to be broadly positive:

- A 2017 meta-analysis[1] summarizing 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, e.g. in class (*g*=0.61, p<.001);
- A 2018 meta-analysis[2] summarizing 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, e.g. first at home and then in class (*d*=0.40, 95% CI [0.31, 0.50]).

According to the above-mentioned 2017 meta-analysis, the effect of ST on learning outcomes seems to be stronger:

- When the initial learning involves reading or studying a passage, rather than listening;
- For mixed-format practice tests (i.e. including a mix of free-recall, cued-recall and short-answer tests) than for practice tests using a single type of test;
- For secondary school students than for students at other levels;
- When the practice and final tests formats are identical, although this point is contested (see below);
- When the time lag between practice and final test is between 1 and 6 days;
- For high treatment fidelity studies.

The same meta-analysis indicates that the effect of ST seems to be broadly similar:

- With or without feedback;
- When the final test was administered in the class room or in a lab.

[1] http://journals.sagepub.com/doi/abs/10.3102/0034654316689306

[2] http://acsweb.ucsd.edu/~scp008/pdf/PR_2018.pdf

However, it is important to bear in mind that ST can have adverse effects on:

- Pupils' *deep* learning: Some experts[3,4,5] have warned that an over-utilisation of tests (including self-tests) in school curricula can result in superficial (or shallow) learning, which, by definition, can't be detected in standard tests.
- Pupils' emotional health and well-being: The World Health Organisation (2012)[6] found that 11 and 16-year-old pupils in England feel more pressured by their school work than is the case in the vast majority of other European countries. McCaleb-Kahan and Wenner (2009)[7], drawing on research in the USA, 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.

The aim of this evaluation is to explore the advantages and disadvantages of self-testing in science among both pupils and teachers.

1.3. Who

The SSTT was developed by the Kingsbridge, Durrington and Huntington Research Schools.

The delivery model is the following:

- At national level, the three research schools will act as local 'hubs';
- At local level, three other schools will be recruited in each hub to help deliver the SSTT, bringing the total number of pilot schools to 12.
- At school level, the intervention will be delivered by science teachers, led by Heads of Department and supported at home by parents.

The intervention is expected to benefit all children in participating schools' Year 10 Science classes.

The intervention will be evaluated by NatCen Social Research.

The EEF funded both the development of the intervention and its evaluation.

1.4. What

The SSTT is made up of five evidence-informed, content free (and thus adaptable) strategies to be used and deployed by teachers, pupils and parents. These five strategies are:

- Pre/Post tests (students completing tests at the start and ends of topics)
- Flash card revision (the key being ensuring best practice use of these so that they do include self-testing and also interleave ideas from different topics)
- Mindmapping tests (students are given blank structures to fill in and then use the blank templates to test themselves)
- Structured note taking (including writing revision questions in the margin of notes that can be used during revision)
- Cumulative quizzing that include questions from not only the current topic, but also those taught earlier in the year

The Head of Science and one other science teacher from each school will attend two days of training at one of three EEF Research Schools (Huntington, Durrington, and Kingsbridge).

[3] http://acme-uk.org/media/10498/raisingthebar.pdf

[4] https://eric.ed.gov/?id=EJ577199

[5] https://link.springer.com/article/10.1007/s10648-013-9248-9

[6] http://www.euro.who.int/__data/assets/pdf_file/0003/163857/Social-determinants-of-health-and-well-being-among-young-people.pdf

[7] https://opencommons.uconn.edu/nera_2009/27/

Each research school will provide training for four schools, who will be the pilot schools. Pilot schools will also receive two half days of in-school coaching to help to tweak their practice and ensure good implementation.

Teachers will be given the generic toolkit and guidance on how to use it and the theoretical basis of the intervention's aim to increase the amount of pupil self-testing in KS4 science study will be explained. The intersession tasks require them to translate tools into practice in the classrooms.

1.5. Evaluation overview

The objectives of this study are to assess the following IPE dimensions:

- **Evidence of promise**, i.e. the extent to which the intervention delivered its main outcomes at the pilot stage. The primary outcome of the pilot is the effect of the intervention on pupils' learning methods and outcomes, as perceived by the pupils and the teachers.

- **Feasibility**, i.e. the sum of all drivers and obstacles to the success of the intervention at the pilot stage. We will compare and contrast the perspectives of:
  - Heads of Science;
  - Teachers; and
  - Parents.

- **Scalability**, i.e. the capability of an intervention to deliver similar outcomes as in the pilot when scaled up. This capability depends on two main factors:
  - The replicability of the intervention in different contexts. For example, if the intervention relies too heavily on the commitment of developers (gold-plating), it might not be replicable on a larger scale.
  - The representativeness of the conditions in which the intervention was piloted. For example, if the intervention was piloted in high-performing schools, there is no guarantee that the results will be similar in average- or low-performing schools.

2. Research questions

IPE Domain	RQ#	Research Questions	Source of Data
Evidence of Promise			
Pupils' perspectives	RQ1	What strategies did pupils employ to study lessons before the intervention?	SD8
	RQ2	How time-consuming were previous strategies to memorise lessons?	SD8
	RQ3	How effective were previous strategies to memorise lessons in the short term? In the long term?	SD8
	RQ4	How effective were previous strategies to understand lessons?	SD8
	RQ5	How often did pupils use ST at home and for how long?	SD8
	RQ6	How often did pupils use ST in class and for how long?	SD8; SD6
	RQ7	To what extent did ST change (i) pupils' study strategies; and (ii) the time spent studying at home?	SD8
	RQ8	To what extent was this dosage sufficient?	SD8
	RQ9	Did pupils get enough support from teachers/parents?	SD8
Pupils' vs. teachers' perspectives	RQ10	Did ST seem to affect retention, comprehension and transfer of tested items?	SD8; SD3; SD6

	RQ11	Did ST seem to affect retention, comprehension and transfer of non-tested items?	SD8; SD3
	RQ12	Which of the 5 activities seemed to be most/least effective?	SD8; SD3; SD6
	RQ13	Was ST equally effective at different retention intervals?	SD8; SD3
	RQ14	Did ST seem to cause stress or other adverse effects?	SD8; SD3; SD6
	RQ15	Was corrective feedback helpful? Or would it be helpful?	SD8; SD3
	RQ16	What changes to the toolkit should be made?	SD8; SD3
Feasibility			
Heads of Science's perspectives	RQ17	Why did HoS decide to participate in the pilot?	SD2
	RQ18	How supportive were headteachers?	SD2
	RQ19	How easy/costly was it to coach fellow teachers?	SD2; SD5; SD10
	RQ20	How easy/costly was it to engage with parents?	SD2; SD7; SD10
	RQ21	Did HoS get sufficient and appropriate training?	SD2; SD4
	RQ22	How easy/costly was it to customise the toolkit?	SD2; SD10
	RQ23	How easy/costly was it collect monitoring data?	SD2; SD10
	RQ24	Will science teams keep using the ST toolkit beyond the pilot phase?	SD2
	RQ25	Will science teams introduce ST at other key stages?	SD2
	RQ26	What changes to the intervention should be made?	SD2
Teachers' perspectives	RQ27	What strategies did teachers employ before the intervention to help pupils memorise lessons?	SD3
	RQ28	How effective were these strategies in terms of retention and understanding?	SD3
	RQ29	Did teachers support and understand the intervention?	SD3; SD5
	RQ30	Did teachers get sufficient and appropriate coaching?	SD3; SD5
	RQ31	Did the twilight sessions bring additional clarity about the project?	SD3
	RQ32	Did the intervention effectively increase the use of ST in class?	SD3
	RQ33	Did the intervention effectively increase the use of ST at home?	SD3
	RQ34	Did all pupils benefit equally from the intervention?	SD3
Parents' perspectives	RQ35	Did parents support and understand the intervention?	SD7
	RQ36	Did parents attend the information session? Did they find it useful?	SD7
	RQ37	Did parents read the project documentation? Did they find it useful?	SD7
	RQ38	Did parents support pupils in using the toolkit?	SD7
	RQ39	What changes to the intervention should be made?	SD7
	RQ40	What evidence is there that parents engaged with text messages, emails, infographics, etc.?	SD9
Scalability			
Developers' perspectives	RQ41	How clear and robust is the LM?	SD1
	RQ42	How easy/costly was it to develop the intervention?	SD1; SD10
	RQ43	How easy/costly was it to recruit pilot schools?	SD1; SD10

	RQ44	How easy/costly was it to train coaches?	SD1; SD10
	RQ45	How easy/costly was it to keep the project on track?	SD1; SD10
	RQ46	What is the perceived level of implementation fidelity?	SD1
	RQ47	What elements of the intervention need to be improved?	SD1
Evaluators' perspectives	RQ48	To what extent are pilot schools representative of the population of UK schools? (provider level data)	SD11
	RQ49	To what extent are participating pupils representative of the population of UK pupils? (pupil-level data)	SD11
	RQ50	Are there any other reasons why the effect of the intervention might be different at national level?	SD11

3. Methods

3.1. Data collection

Our evaluation will be based on 11 sources of data (SD), which are presented below. All sampling methods are detailed in section 3.3 below.

Individual and group interviews

SD1 We will conduct two **Developer Focus Groups (DFG)** to clarify and evaluate the Logic Model (LM):

- The pre-intervention DFG will be used to clarify the LM and will be run in accordance with EEF guidance (IDEA workshop).
- The post-intervention DFG will be used to evaluate the LM based on the experience of developers.

SD2 We will conduct two rounds of **semi-structured interviews with the Heads of Science** in each sampled school to (1) understand the motivations of those ‘championing’ the project in schools; (2) understand the barriers and facilitators to delivery at school level; and (3) gather information about the setting to inform sampling of the schools.

- Early implementation interviews conducted after the training workshop will 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 will explore the feasibility and acceptability of the ST toolkit, barriers and facilitators to delivery, perceived impact for professional practice and pupil outcomes. The interviews will also explore recommendations for improvement to the ST toolkit and to the support interventions (e.g. training, parent engagement, etc.).

SD3 We will conduct two rounds of **semi-structured interviews with sampled teachers** to assess (1) the acceptability of ST among participating teachers; (2) the perceived effect of ST on pupils’ learning; and (3) the effect of ST on teaching practices.

- Early implementation interviews conducted just after the first coaching session will assess the acceptability of ST among teachers (both for pupils and for themselves), explore teachers’ strategies to boost retention and understanding (pre-intervention) and evaluate the quality of the coaching session.
- Post-interventions interviews will be used to assess the perceived effect of the intervention on pupils’ retention, understanding and well-being, as well as on teaching practices.

Observation of key events

SD4 We will **observe all training workshops** (i.e. both day 1 and day 2 in each hub) to assess how information was cascaded from hubs to heads of science.

SD5 We will observe one **in-school coaching session** per sampled school to assess how information was cascaded from heads of science to other teachers.

SD6 We will observe one **science class per sampled teacher** to understand (1) how information was cascaded from teachers to pupils; (2) the dosage and fidelity of implementation at class level; (3) teachers' and pupils' engagement with the toolkit.

SD7 We will observe three **parent info sessions (one per hub)** to assess (1) how information was cascaded from schools to parents; (2) parents' understanding of the benefits of ST; and (3) the acceptability of the toolkit.

Surveys

SD8 We will conduct a **pupil online survey** to (1) assess the perceived effect of the toolkit on learning and well-being outcomes; and (2) estimate the response rate, should the intervention be trialed. Given the natural cognitive development of pupils over the course of the year, the intervention is unlikely to be the only factor that could affect memorization and learning strategies. Thus, we propose to survey pupils only once, in the middle of summer term 2019. To measure the perceived effect of ST, we will ask pupils to compare their learning strategies and outcomes before and after the introduction of the toolkit. The survey questionnaire will be short (20 minutes maximum). We recommend that teachers administer the survey in class to maximize the response rate. A high response rate will allow us to run a few subgroup analyses.

Administrative data

SD9 We will collect any **app and/or usage data** that will be generated during the project from any digital tool (ST toolkit, parent toolkit). Data might include the number of single-user visits, the time spent using the tool, etc. This will allow us to assess pupils' and parents' engagement with ST.

SD10 We will collect **cost data** to assess the affordability of the intervention. We will provide Developers with a pro-forma to estimate the costs of developing the ST toolkit and providing training/support for schools. Additionally, the post-survey will ask Heads of Science to estimate the demand of the intervention on staff time, and any other costs associated with delivering the ST toolkit, including any hardware.

SD11 We will conduct **desk research** to assess the extent to which pilot schools are representative of the population of UK schools and whether similar effects can be reasonably expected after scale-up. We will compare pilot schools with the average school in the UK using: OFSTED rating, GCSE results, class size, proportion of FSM students, proportion of BME students, etc.

3.2. Recruitment

Unit of analysis	Number	Sampled	Rationale
Organisations			
School hubs	3	3	All school hubs will be included in the evaluation.
Participating schools	12	6	We will select two schools in each of the three hubs. These schools will be purposively selected to provide range and

			variation, including with regard to: size, existing practice with regard to ST and GSCE results. The sampled schools will be used as case studies.
		6	SD9: Assuming the engagement of parents can be monitored, we will collect monitoring data from each of the six sampled schools.
		6	SD10: We will provide Heads of Science in each of the six sampled school with a pro-forma to help them assess the cost of the intervention at school level.
		12	SD11: We will conduct desk research for all participating schools to determine how representative they are of all English schools.
People			
Developers	3	3	SD1: The Developer Focus Groups will involve a developer from each of the 3 hubs.
Heads of Science	12	6	SD2: We will interview the Head of Science in each of the six sampled schools.
Teachers	36	6	SD3: We will interview one teacher in each of the six sampled schools. These teachers will be selected with a view to provide a range of views with regard to the acceptability of ST and current teaching practice. This will be based on information provided by the Heads of Science.
Pupils	Unknown	All	SD8: We will survey all participating pupils across all participating schools. The aim is to maximise sample size.
Parents	Unknown	Unknown	SD7: We will gather feedback from parents participating in the information sessions in each sampled school. Feedback forms will be distributed at the start of the session.
Events			
Training sessions	6[8]	6	SD4: All training sessions will be observed (two sessions in each of the three hubs). We expect interesting variations in terms of engagement and feedback from the first to second session.
In-school coaching sessions	24[9]	6	SD5: We will observe one coaching session in each of the six sampled schools. We will 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 will observe one science class for each of the 6 sampled teachers.
Parent info sessions	12[10]	3	SD7: We will observe one parent information session in three of the six sampled schools (and one in each of the three regions). We will seek to obtain a range of locations (urban/rural) and school performances based

[8] 2 training days x 3 school hubs

[9] 2 coaching sessions x 12 participating schools

[10] 1 information session x 12 participating schools

	on information provided by Heads of Science and our own research.

3.3. Data analysis

Raw, qualitative data will be analysed thematically, using the Framework approach. This will allow us to analyse the data by theme and by case. For example, we will analyse barriers and facilitators across all case studies, identifying similarities and differences by school type. Within-case analysis will enable us to triangulate the perspectives of providers and developers across the year in order to come to a holistic picture of pilot implementation.

Raw, quantitative data will be analysed by means of frequencies and cross tabulations, using SPSS.

4. Ethics and registration

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 guidance and recommendations provided by the REC have been incorporated in this study plan.

5. Data protection

NatCen is the data controller and processor for this evaluation.

5.1. Personal data

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

5.2. Special data

We will not process special categories of data as part of this study.

5.3. Assessment data

We will not process assessment data as part of this study.

5.4. Data processing

NatCen will provide a Memorandum of Understanding to participating schools, explaining the nature of the data being requested, how it will be collected, and how it will be passed to and shared.

Procedures for ensuring data quality, anonymity and confidentiality can be found in our privacy notice, which is available here: http://natcen.ac.uk/help/privacy/.

6. Personnel

6.1. Delivery team

The Delivery team includes:

- **Lorwyn Randall** (Kingsbridge Research School); Strategic Lead – overseeing programme design, resource development, workshop delivery and follow-on support.
    - **Jon Eaton** (Kingsbridge Research School); Project Lead – Workshop facilitation and coaching support
- **Jane Elsworth** (Huntington Research School); Strategic Lead – programme design team, resource development, workshop delivery and follow-on support.
    - **Penny Holland** (Huntington Research School); Project Lead – Workshop facilitation and coaching support
- **Shaun** Allison (Durrington Research School); Strategic Lead – programme design team, resource development, workshop delivery and follow-on support.
    - **Steph Temple** (Durrington Research School); Project Lead – Workshop facilitation and coaching support

6.2. Evaluation team[11]

Conceptualisation	AV
Data curation	SF; MM; HB
Analysis	AV; SF; MM; HB
Funding acquisition	AV
Investigation	SF; MM; HB; BT
Methodology	AV
Project administration	AV; SF
Resources	NatCen Social Research
Software	SF
Supervision	AV
Validation	AV; SF
Visualisation	SF; MM; HB
Writing – original draft	AV; SF; MM; HB
Writing – review and editing	AV; SF

AV: Arnaud Vaganay (Principal Investigator);
SF: Sarah Frankenburg (Senior Researcher);
HB: Helen Burridge (Researcher);
MM: Molly Mayer (Researcher);
BT: Bethany Thompson (Research Assistant).

All evaluators are affiliated with NatCen Social Research.

7. Risks

Risk	Likelihood / Impact	Mitigation/Contingency

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

Difficulty scheduling school visits within the required timescale	**Likelihood:** Low **Impact:** Medium	The initial recruitment materials will set out clearly the data collection points and details about the pilot study so that schools are making an informed decision about whether to participate. Sufficient resource allocated to arranging visits and large team means we have flexibility.
Not possible to schedule all the necessary data encounters during the school visit.	**Likelihood:** Medium **Impact:** Medium	Clarity about expectations of pilot participation and flexibility in scheduling visits will mitigate this risk. We will supplement the face-to-face data encounters with follow-up phone calls.
Lack of engagement from staff with the self-testing approach undermines ability of pilot to assess IPE dimensions	**Likelihood:** Low **Impact:** High	Recruitment materials will emphasis importance of buy-in from classroom staff. Evidence base of intervention should help secure teacher engagement. Extent of engagement with the ST approach will be investigated as part of the pilot and recommendations made for scale-up.
Pupils do not complete the survey	**Likelihood:** Low **Impact:** Medium	We propose that the survey will be completed during school time. The survey will be short.
Schools drop out of the pilot	**Likelihood:** Low **Impact:** Low	This is a formative study mainly using qualitative methods. We will give schools advance notice of research activities and arrange interviews and visits to suit availability and the school day. We will work closely with Kingsbridge Academy staff to address any school concerns about research burden.

8. Timeline

Research tasks	Start	Finish
Set-up meeting	21/09/18	21/09/18
SE1: 1st Developer FG	03/12/18	03/12/18
MS1: Submission draft study plan	**31/01/19**	**31/01/19**
SE2: 1st interview - Heads of Science	20/01/19	10/02/19
SE4: Observation of training workshop 1	20/01/19	07/02/19
SE5: Observation of coaching session 1	01/02/19	20/02/19
SE3: 1st teacher interview	01/02/19	20/02/19
SE4: Observation of training workshop 2	01/03/19	14/03/19
SE7: Observation of parent info sessions	01/03/19	17/05/19
SE6: Observation of science classes	01/03/19	17/05/19
SE3: 2nd teacher interview	01/03/19	17/05/19
MS2: Completion of observations	**31/05/19**	**31/05/19**
SE8: Pupil survey	15/05/19	15/06/19
SE2: 2nd interview - Heads of Science	15/05/19	15/06/19
MS3: Completion of survey + interviews	**28/06/19**	**28/06/19**

SE1: 2nd Developer FG	01/07/19	15/07/19
SE10: Cost data analysis	16/07/19	31/07/19
SE9: App/usage data	16/07/19	31/07/19
SE11: Desk research	15/08/19	31/08/19
Data triangulation/ synthesis	01/08/19	13/09/19
MS4: Slidepack, presentation, spend	**20/09/19**	**20/09/19**
Draft report	20/09/19	11/10/19
MS5: Draft report	**25/10/19**	**25/10/19**
Final report	01/11/18	30/11/19
Submission of data to EEF archive	01/12/19	13/12/19
Final statement of spend	01/12/19	13/12/19
MS6: Final report, archiving, spend	**27/12/19**	**27/12/19**