Education Endowment Foundation

# Let's Think Secondary Science

## Evaluation report and executive summary

July 2016

**Independent evaluators:**

Pam Hanley, Jan R Böhnke, Bob Slavin, Louise Elliott and Tim Croudace

UNIVERSITY of York

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.

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

**Danielle Mason**
Head of Research and Publications
Education Endowment Foundation
9th Floor, Millbank Tower
21–24 Millbank
SW1P 4QP

**p**: 020 7802 1679
**e**: danielle.mason@eefoundation.org.uk
**w**: www.educationendowmentfoundation.org.uk

## About the evaluator

The project was independently evaluated by a team from the Institute for Effective Education (IEE), University of York: Pam Hanley, Bob Slavin, and Louise Elliott; and from the Hull York Medical School and Department of Health Sciences, University of York: Jan R Böhnke and Tim Croudace.

The lead evaluator was Pam Hanley.

Contact details:

Dr. Pam Hanley

**Institute for Effective Education**
University of York
Heslington, UK
YO10 5DD
Tel:  01904 328166
Email: pam_hanley@hotmail.com

# Contents

# Executive summary

## The project

Let’s Think Secondary Science (LTSS) aims to develop students’ scientific reasoning by teaching them scientific principles such as categorisation, probability, and experimentation. LTSS challenges students’ thinking, develops their metacognitive skills, and encourages cooperative learning. Let’s Think Forum (a registered charity led by a group of academics, teachers, and consultants) created LTSS by adapting a programme called Cognitive Acceleration through Science Education (CASE). Let’s Think Forum’s adaptations to CASE included reducing the number of lessons from 30 to 19, reducing the number of scientific principles that are taught, rewriting lesson plans, and providing additional resources like PowerPoint slides and video tutorials. In this trial, Let’s Think Forum provided one day of training and three support sessions per year to the science teachers who would be teaching LTSS. These teachers then delivered the lessons to a cohort of Year 7 students instead of their usual science lessons. LTSS was delivered over two years, so this cohort of pupils continued to receive the programme in Year 8.

LTSS was evaluated using a randomised controlled trial with over 8000 students in 53 schools. Schools were randomly allocated to deliver either the programme or their ‘business as usual’ science teaching. It should be considered an effectiveness trial, as it aimed to test a scalable intervention under realistic conditions in a large number of schools. The primary outcome measure was an age-appropriate science test based on a Key Stage 3 SATs paper, and the secondary measures were English and maths tests provided by GL Assessment. The process evaluation consisted of lesson observations, surveys, and interviews with staff, and surveys and focus groups with students. The trial started in September 2013 and ended in July 2015.

### Key conclusions

1. This evaluation provided no evidence that Let’s Think Secondary Science improved the science attainment of students by the end of Year 8.
2. Students who received LTSS did worse than the control group on the English and maths assessments, but this result could have occurred by chance and we are not able to conclude that it was caused by the programme.
3. Many schools did not implement the programme as intended by the developer. In many schools, individual teachers delivered fewer than the full programme of 19 lessons and senior leaders were less engaged than prescribed by the programme.
4. Although most teachers were providing opportunities for students to work collaboratively, there was some evidence that more support to help teachers promote effective small group discussions would be welcomed.
5. Previous evaluations of CASE have suggested that it had longer-term impacts on academic attainment. Future research could examine whether LTSS also has a long-term impact by examining the GCSE results of the pupils involved in this evaluation.

## How secure are the findings?

***Security rating awarded as part of the EEF peer review process***

Overall, the findings from this evaluation are judged to be of moderate security. It was a large and well-designed randomised controlled trial. At the beginning of the trial, the schools and pupils who received the intervention were similar to the schools and pupils in the comparison group. Two padlocks were removed from the rating because 26% of the pupils did not complete all the required tests and were not included in the final analysis. 11% of schools dropped out of the study. There were no other major threats to the security of the trial.

## What are the findings?

This evaluation provided no evidence that LTSS had an impact on science attainment. Students who received LTSS did worse than the control group on the English and maths assessments, and girls who received LTSS did slightly better than the control group in the science assessments, but these results could have occurred by chance and it is not possible to conclude that they were caused by the programme. However, they could provide useful areas of focus for future evaluations of LTSS. There was no evidence that the programme had a differential impact on pupils according to their eligibility for free school meals (FSM), or prior attainment in Key Stage 2 (KS2) maths and reading.

The process evaluation suggested several possible reasons why the programme did not have an impact. Schools often did not implement the programme as intended. Although just over half the teachers had taught at least 13 of the 19 different LTSS lessons, a quarter of them had taught 6 or fewer. This did not necessarily mean that students would not receive all 19 lessons, since the survey was completed before the end of the year and also the teachers could have been sharing delivery with a colleague. However, when teachers did teach an LTSS lesson, they generally appeared to do so as the programme developers intended. A member of each school’s leadership team was supposed to oversee the school’s implementation of the programme, but senior leaders appeared to be involved to a much lesser degree than expected, if they were involved at all. LTSS encouraged teachers to deliver each lesson first to the year group above the trial cohort, so that they would have the opportunity to practise delivery. Only two in five of the teachers reported that they did this very or quite often, and a quarter reported that they never did. Teachers also reported that it was challenging to fit all the lesson material, which was designed for hour-long lessons, into the 50-minute slots that many schools used.

Although the large amount of group work was popular with most students, some teachers complained about it leading to disruptive behaviour. This suggests that some teachers needed more support to be able to encourage more productive small group discussion. Some teachers felt that LTSS was less accessible for lower-attaining students. However, it should be noted that the impact evaluation did not find any evidence that lower-attaining students who received the intervention performed worse than similar pupils in the control group. Student reaction was mixed, with many enjoying the discursive and collaborative nature of LTSS, but some students held negative views about the repetitiveness of lessons and working in groups. Overall, most teachers enjoyed the challenge of LTSS, were positive about the practical activities, and recognised improvements in their questioning techniques.

Previous evaluation of CASE also failed to detect an impact on science attainment immediately after the intervention finished at the end of Year 8. However, there is evidence that CASE had a long-term impact on attainment, as students who received CASE outperformed the control group in their GCSE science, maths, and English language exams. Future research could examine whether LTSS also has a long-term impact by examining the GCSE results of the pupils involved in this evaluation.

## How much does it cost?

The overall cost per school for the two-year programme was £4,490. The average cost per student is estimated as £3.99 per year across three years. The main costs related to the training, with much lower amounts for lesson resources and equipment. LTSS involved some burden of extra time, which the delivery team estimated at five hours per year for teachers (mainly planning) and two hours per year for a technician.

| Group | Effect size (95% Confidence interval) | Estimated months’ progress | Security rating | Cost rating |
|---|---|---|---|---|
| **LTSS versus control group** | -0.02<br>(-0.09 to 0.04) | -1[1] | 🔒🔒🔒🔓🔓 | **£** |
| **LTSS FSM versus control group** | -0.03<br>(-0.18 to 0.07) | -1 | | |

[1] *Since this report was published, the conversion from effect size into months of additional progress has been slightly revised. If these results were reported using the new conversion, all measures would be reported as 0 months of additional progress rather than -1. See* ***here*** *for more details.*

# Introduction

## Intervention

Let’s Think Secondary Science (LTSS) is based on Cognitive Acceleration through Science Education (CASE), a programme that was first developed and evaluated in the 1980s. Like CASE, LTSS was designed to promote better thinking by providing students with cognitive challenge in the context of science education. It reflects the social construction of knowledge by promoting collaborative working, and encourages students to reflect on their own thinking and learning (metacognition).

CASE was based on teachers delivering an hour-long session every fortnight over two years for a total of 30 lessons. The LTSS developers considered that some of this material could be merged or dropped, and the full LTSS programme consisted of 19 one-hour lessons. Like CASE, it was designed to be delivered by Science teachers across the first two years of secondary school (Years 7 and 8). Although it was intended to be delivered during time set aside for science, the lessons were not designed to directly address individual elements of the science curriculum. Rather, the aim was that learners would engage emotionally and cognitively; would construct meaning collaboratively through group work; and would reflect on the new ideas over time (before, during and after the lesson) and across contexts (linking for instance to previous learning, or understanding how the new learning could be transferred to other situations).

Significant similarities between CASE and LTSS are that they both:

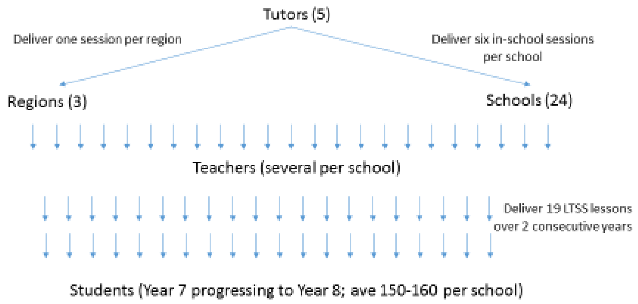
- aim to improve reasoning;
- aim to include the same level of challenge in any lesson;
- introduce pupils to the same scientific reasoning concepts (where they had not been omitted, see differences below); and
- are based on the same pedagogical principles.

Significant differences included the following:

- LTSS had fewer lessons: 19 as opposed to 30.
- LTSS omitted some scientific reasoning concepts that were introduced in CASE.
- LTSS had fewer lessons per scientific reasoning concept.
- LTSS offered different support and materials to teachers: simplified pedagogical principles, fewer support notes, video tutorials, peer co-planning.
- LTSS offered different materials and presentation to students: digital resources, cartoon worksheets, and more group work.
- LTSS involved CPD with less tutor contact and more in-school responsibility.
- LTSS required that teachers deliver each lesson twice to different classes, so that they had the opportunity to practise.

The intervention focused on a single cohort of students, who started in Year 7 and were followed to the end of Year 8. In terms of teacher training, there was one day per year for each region to meet together as a cluster of schools. The rest of the support was school-based, with each school having three days of support in year 1 and three in year 2 (six in total). The visits were agreed with the school and were a mixture of sessions (half days, evening sessions, Skypes, etc). This flexibility was intended to provide responsive support to meet the needs of individual schools. The aim was that the tutor built up a long-term personal relationship with each school and science department (see Figure 1). In total across the three geographic regions of the trial (north-east, midlands and south-west England), a team of seven tutors was involved in delivering the LTSS programme to the teachers.

Figure 1: LTSS delivery model (2 years)


Tutors (5)
Deliver one session per region
Deliver six in-school sessions per school
Regions (3)
Schools (24)
Teachers (several per school)
Deliver 19 LTSS lessons over 2 consecutive years
Students (Year 7 progressing to Year 8; ave 150-160 per school)

Each teacher was asked to teach each LTSS lesson twice (the first iteration being delivered to a class from the year group above the study cohort) so that lesson delivery could benefit from a practice effect on the second occasion. Schools were provided with lesson plans and resources for many of the lessons, for instance sort cards and PowerPoint slides.

Lessons tend to follow a similar overall pattern. They start with an up-to-date hook to engage students and provide them with concrete examples and materials. Students are then encouraged to think about a series of progressively more challenging problems, working collaboratively with peers to promote exploratory talk and social construction of knowledge. Towards the end of the session the teacher encourages them to reflect on their learning and to broaden their focus from the lesson specifics to other contexts.

The outline for the seventh LTSS lesson provides an illustration of the increasing cognitive challenge within sessions. The topic is classification and it builds directly on the previous lesson. The aims are that students:

- realise that grouping can be based on different criteria (e.g. where animals live or what they eat);
- understand that some criteria are more robust because they stay consistent over time or other people can use them to get the same result; and
- actively engage in classifying by setting boundaries for categories of a continuous variable such as size.

The session starts with the 'hook' of showing students an image of different technologies to promote reflection and discussion of ideas covered in the previous LTSS lesson.

As a first activity, students are shown an image of animals around a pond and asked a series of questions, such as how they would group the animals, and are there more animals that are not ducks or more animals that are not birds (to promote thinking in the negative and consideration of sub-sets). In the second activity, students are asked to come up with four possible characteristics to sort birds. They are then asked to group sort-cards of birds by size, so they experience the difficulty of classifying by a continuous variable. They are asked to regroup the birds based on a more useful characteristic. Finally, they are given a card with a humming bird and asked if they can fit it into their system.

As homework, students are encouraged to identify two sets of things that they or people they know find it useful to classify (unrelated to science lessons) and to design a classification system for one of these.

As part of the LTSS model, each school is encouraged to take ownership of the approach and allocates a member of the senior leadership team (not necessarily a science specialist) to oversee the project. The senior leader is expected to conduct lesson observations, including one shared with the LTSS tutor in year 1. The tutor and senior leader support the development of an in-school learning group whereby the teachers can share responsibility for joint lesson planning, peer coaching, and considering student feedback. Teachers receive feedback on their adoption of the LTSS approach as well as having tools for self-reflection such as a journal, checklist, and summary of how teacher expertise progresses. Individual support is tailored to the teacher’s need.

There were various issues with the delivery of the intervention. In some schools, only a few teachers or classes used LTSS—in three identified cases, only part of the cohort received the LTSS intervention (i.e. some students within the intervention school received the control ‘business as usual’ condition) and in another case, the majority of students only received LTSS in Year 8. In other schools, LTSS was taught to the entire cohort as intended but not all students received all 19 lessons. The level of involvement of a senior leader and the frequency of lesson observations also varied. The process evaluation also revealed challenges around differentiation and successful small group work.

## Background evidence

The two-year CASE programme on which LTSS is based has been evaluated over the last 20 years with positive indications across outcomes including scores on cognitive development measures (Piagetian reasoning tasks) and better-than-expected gains in GCSE maths and English as well as science.

The first published study was conducted in the 1980s by the programme originators, Adey and Shayer (1993). This found an immediate positive impact on CASE students’ cognitive development as measured by the Piagetian tests; a delayed effect on science attainment (picked up at GCSE); and a transfer effect to English and maths (also based on GCSE results). However, the sample was small (24 classes, split between control and intervention, in 9 schools) and the immediate pre/post-test measures were based on the same model of Piagetian reasoning as was the intervention, making the tests inherent to the treatment. Furthermore, the analysis ignored the two schools (representing two intervention and two control classes) that dropped out or failed to implement the intervention. This could lead to a false positive bias if those two discounted intervention classes had achieved markedly worse scores than the remainder of the intervention group.

Shayer (2000) reports positive effects on GCSE science, English and maths results in a later study with a quasi-experimental design looking at 11 CASE schools and 16 matched controls. Although the sample of schools was larger, the rationale behind the selection of the schools is unclear.

Favourable evaluations of CASE have been conducted by other researchers in several countries, although all have design limitations. Studies designed to show the impact on cognitive development have been positive (Iqbal & Shayer, 2000; Choi et al, 2002; McCormack et al, 2014) but all used Piagetian-based tasks as the outcome measure. This may give an in-built advantage to CASE participants as it is a Piagetian-inspired programme.

Among the studies that included an attainment-based outcome measure, some had weak comparison groups (e.g. Endler & Bond, 2001; Jones & Gott, 1998; Oliver et al, 2012). Several studies had low sample sizes (Mbano, 2003), including those conducted in only one school (Babai & Levit-Dori, 2009; Maume & Matthews, 2000).

The favourable but inconclusive evidence around cognitive gains and delayed effects on science, English and maths attainment for CASE provides a promising basis for LTSS. The LTSS model also focuses on two approaches identified as having high evidence of impact in the EEF’s Teaching and Learning Toolkit: collaborative learning, and metacognition and self-regulation (sometimes called ‘learning to learn’ approaches). A more robust evaluation of LTSS therefore seemed justified.

The evaluation reported here was set up as an effectiveness trial, to test a scaled-up version of LTSS with randomly assigned participating schools and rigorous outcome measures.

## Evaluation objectives

It was originally intended that the evaluation would explore the following questions:

- What is the impact of Let’s Think Secondary Science on student achievement in:
  - science
  - maths
  - English?
- What are the effects on the development of cognitive reasoning?
- Are any gains in cognitive reasoning correlated to gains in student achievement?

Questions 2 and 3 were not pursued due to the lack of a suitably robust outcome measure. The original intention was to use GL Cognitive Abilities Test (CAT4). However, the publishers informed us that the test was not designed to assess the sort of changes in capabilities that would be influenced by an intervention, meaning that a lack of effect would therefore not be conclusive. This, along with its burdensome nature as a test (requiring up to three lesson slots for administration), led to it being dropped. The original measures of cognitive development used by Adey and Shayer (1993) were not considered suitable because they could be biased towards the LTSS students for the reasons discussed in the previous section. No other appropriate measure could be identified.

## Project team

On behalf of the Let’s Think forum, the LTSS programme was developed by John Crossland and Barry Gunter. It was delivered by a team comprising Alan Edmiston, Stuart Twiss, Anita Backhouse, Lorraine McCormack, David Bailey, Julian Clarke, and Kate Donegan.

Pam Hanley led the University of York evaluation team which consisted of Robert Slavin (co-investigator), Louise Elliott (data manager), Tim J. Croudace (psychometrician), Jan R. Böhnke (evaluator and research methodologist), and project assistants Imogen Fountain, Kate Thorley, and Sarah Hogben.

## Ethical review

The evaluation team obtained ethical approval from the Department of Education, University of York Ethical Review Panel on 29 May 2013. Headteachers signed an agreement outlining the main commitments of the three parties in the study: the school, the project developers, and the evaluators. The evaluation team provided information and opt-out consent forms for parents/guardians.

Data was managed in accordance with the Data Protection Act (1998). The trial database is securely held and maintained on the University of York’s research data protection server, with non-identifiable data. Confidentiality is maintained and no one outside the trial team has access to the database. Data was checked for missing elements and/or double entries. All outputs were anonymised so that no schools or students could be identified in any report or dissemination of results.

## Trial registration

This trial was registered at http://www.controlled-trials.com/ISRCTN08354937

# Methods

## Trial design

The evaluation was designed as a two-armed clustered randomised controlled trial (RCT). School-level randomisation was chosen in preference to within-school randomisation because teacher collaboration is inherent to the intervention especially in terms of joint lesson planning sessions. Within-school randomisation would have carried with it a high risk of diffusion between experimental and control teachers, as well as potentially reducing the effectiveness of the intervention. Schools allocated to control were asked to continue delivering their normal science lessons (i.e. business as usual) but were placed on a waitlist for the intervention to reduce the chance of teachers becoming demoralised post-allocation and to keep them engaged during the trial period. They were due to start LTSS from September 2015.

## Outcome measures

All outcome measures were administered in June/July 2015.

The primary outcome at the end of the study was science attainment, measured using one paper of the science Key Stage 3 SATs (2009 Tier 3-6 Paper 2). The KS3 science SATs test was designed to be administered to students towards the end of Year 9, when they were due to have completed the KS3 curriculum. After 2008, the tests were no longer compulsory and the 2009 paper was the last one published. Nonetheless, at the time of testing it was still the most appropriate test to use since it was linked to the current science curriculum. The test was designed to cover the full range of content, including scientific enquiry, life processes, materials, and physical processes. Because this test was being administered at the end of Year 8, when some schools would only be part-way through the KS3 curriculum (schools usually take 2–3 years to complete KS3) we administered the easier form of the test (Tier 3-6). There were no signs of a floor effect, i.e. that the test was too hard for some students to access (see Table 4 and Figure 1 in the Technical Appendix). To minimise the burden, only one of the two papers was used. Paper 2 was chosen because it had a slightly wider range of content. Paper copies of the test were used, and students were given 45 minutes to complete it.

Secondary outcomes were measures of English and maths attainment. These were measured by GL Progress Test in English 13 and GL Progress in Maths 13 respectively. Half the schools were asked to administer the English test and half the maths test. This was randomised by school within treatment. The original intention was to administer both these tests digitally, but some schools had problems accessing online versions and had to complete paper copies instead. This slightly delayed the return of the paper tests compared with the online versions. Since the schools affected were reasonably well balanced between intervention and control, there was not considered to be a differential effect by treatment group. Performance on PTE did not seem to be affected by whether administration was online or on paper (Table 13, Technical Appendix), but there was some effect for PIM (Table 14, Technical Appendix) so we controlled for this effect in the statistical model. In terms of content, the online and paper versions are comparable, except that PTE has one extra question on the online compared with the paper version.

The tests were administered by teachers under exam conditions. The science tests were returned to the IEE and scored by trained markers using the national curriculum assessment mark scheme. The English and maths tests were scored by GL, either by computer (if digital versions were completed) or by hand. In all cases, markers were blind to treatment.

In order to minimise the costs and disruptions of data collection, routinely collected English and maths KS2 scores (obtained from the National Pupil Database) were used as the pre-test. There is no longer a mandatory KS2 science test, so the reading and maths scores were used instead. Although they have less predictive power for a science outcome measure, as standardised measures that constitute the main indicators of primary school pupils' academic performance, they are high in contextual validity.

There were some changes to the original protocol. We did not use GL Cognitive Abilities Test (CAT4) as intended because (a) the publishers suggested that it measured abilities that would be resistant to interventions; and (b) it was so burdensome to administer that schools would have been reluctant to schedule the necessary time. This meant that the original intention to adopt a matrix sampling model and use a planned missing data analysis was not pursued.

To make test administration more straightforward for schools, the maths and English tests were each completed by half the schools, rather than half the students within each school as specified in the protocol. Twelve schools had problems accessing computing facilities to complete the GL assessments online as intended, and arrangements had to be made to provide paper copies of the tests instead in these cases. Although the protocol stated we would use English as an additional language (EAL) as a variable in the analysis, we were unable to obtain this data from the National Pupil Database since we had only collected parental opt-out, not opt-in, consent.

## Participant selection

Eligible schools were secondary schools with Year 7 and Year 8 classes (so the study cohort could be followed from Year 7 to Year 8 across the two years of the study). Eligible teachers were those with no previous experience of CASE in the last ten years. It was anticipated that initially all teachers of science to Year 7 (the trial cohort) would be trained and that the training would include teaching assistants if applicable and possible. The project team was responsible for recruiting appropriate schools to the project. A quarter of the sample (12) had a higher proportion of Ever-FSM students than the national average, taken as a proxy for higher poverty catchment areas. To facilitate training and running of the intervention, schools were recruited and randomised in geographical clusters (north-east, midlands and north-west, and south-west).

Memorandums of Understanding were completed by the headteachers of each school before randomisation (see Appendix). Parental consent was sought for students to be involved in the post-testing and potentially to complete surveys and take part in discussions with a researcher towards the end of the two-year trial (see Appendix). No cases of consent being withheld were reported to the evaluators. Schools were asked to provide student UPNs during the second term of the first year of the trial, so in most cases this information was provided over a year before the post-tests were completed. This time lapse was one factor contributing to the student-level attrition.

## Sample size

The sample size was determined using Optimal Design software. The assumptions, based on previous experience of similar studies, were as follows:

Pupils per school per year group: 150
Proportion of variance in the outcome explained by covariates (R-squared): +0.563
Intra-class correlation: 0.12
Criterion for statistical significance (α): p<.05
MDES: 0.20
Power: 0.80

Based on these assumptions it was calculated that a sample size of 50 schools would be needed (25 per treatment group) to detect an effect size of 0.20 (Table 1). The recruitment target was 54 schools

(18 per region) to allow for some dropout. The total number of schools that signed up to the study was 53 (18 in two regions and 17 in the third).

## Randomisation

Schools were randomised within each region except for one recruit to the north-east which signed the MOU too late to be randomised with the other NE schools, so was randomised with the midlands/north-west region. Randomisation used matched pairs of schools stratified (in order of priority) by average percentage GCSE (A*–C) results over three years, percentage of students eligible for free school meals, percentage EAL, and number of students registered at the school. Random number generation was used to identify which school in each pair was the intervention school and which the control. The one unpaired school was allocated to treatment by coin toss. Randomisation was carried out by the Data Manager as follows, with schools assigned codes rather than names to maintain anonymity during the process:

1 July 2013 – 18 SW schools

10 July 2013 – 18 NE schools

12 July 2013 - 16 Midlands/NW schools plus 1 NE school

Although students at all the schools were asked to complete the science attainment outcome measure, schools were only asked to administer one of the secondary outcomes (maths or English). The schools were allocated to PIM or PTE at random on 5 May 2015 by the project statistician. Within each region, half the control schools were randomised to PIM and half to PTE. The intervention school within the pair was allocated to the other test, ensuring that within each pair both tests were used. Where there were unpaired schools, for instance because the other school had withdrawn from the trial, the tests were assigned individually at random (by single draw from a binomial distribution).

## Analysis

An intent-to-treat design was used, meaning that schools were asked to provide evaluation data if they withdrew from the programme after randomisation. To account for clustering, multilevel modelling was used, with students nested within schools and school means compared. The test scores for schools randomly assigned to the Let’s Think programme were compared to those in the randomly assigned control group, controlling for KS2 pre-test scores.

Using the between school fixed effect estimate $\gamma_{01}$ as a measure of absolute controlled difference (controlled for gender and KS2 results), the effect size was calculated using the following formula according to EEF guidelines to approximate Hedge's g for a multilevel intervention study (see details in the Technical Appendix):

$$ES = \frac{\left(\overline{Y}_T - \overline{Y}_C\right)_{adjusted}}{\sqrt{\sigma^2{}_s + \sigma^2{}_{error}}} = \frac{\gamma_{01}}{\sqrt{e_{ij} + u_{00}}}$$

A sensitivity analysis was run by re-estimating the multilevel model on the same sample, but weighting students for their probability of having no observation on the follow-up assessment.

After the main analyses including all students, subgroup analyses were carried out for recipients of free school meals (Ever-FSM), boys and girls, and high, average and low attainment (based on average scores on fine grade reading and maths at KS2).

In addition to the analysis of the primary outcome, the PIM and PTE were collected. To investigate whether differential and/or positive transfer effects can be detected across the different domains that these tests assess, the original plan was to extend the multilevel model from the Primary Outcome

Analysis. Since the overlapping assessment plan originally planned for in this study could not be realised (see above), it was decided later on (based on feedback from the EEF) that only the schools would be analysed that received either the PIM or the PTE in separate sets of analyses. Although this came with disadvantages, particularly in reducing the sample size for the follow-up analysis and thereby reducing the statistical power, this seemed to be the most appropriate strategy.

The same statistical model was run as for the primary outcome, but in addition it was controlled for on school level as to whether the PIM and PTE were administered online or as paper-pencil versions. Again, a positive effect of the LTSS intervention on both of these outcomes would be seen as further corroboration that the LTSS intervention had a positive effect on science attainment and in addition positive transfer effects on mathematical and English attainment could be observed. This analysis would also allow a check on whether there was any detrimental impact on these other subject areas.

A copy of the analysis plan is included in the Appendix, along with technical details of the statistical analyses that were run.

## Implementation and process evaluation

The process evaluation comprised a mix of approaches, including:

- observation of a regional training day;
- case study visits to six LTSS schools: eight lesson observations; LTSS teacher interview/focus group (minimum one per school); one student focus group per school; senior leadership interview (achieved in three schools);
- online LTSS teacher questionnaires: 48 teachers in 17 intervention schools;
- online control teacher questionnaires: 58 teachers in 14 schools (NB including 10 teachers from one school); and
- online student questionnaires: 652 students in 11 intervention schools (excludes 38 who responded to the survey but claimed they had not attended any LTSS lessons).

The link to the online questionnaire was sent to all 21 LTSS schools which were still active in the intervention and (for the teacher survey) to all 27 control schools. Where possible, the intervention and control survey included similar questions, allowing a comparison of teachers’ attitudes and approaches.

All the process evaluation measures were undertaken by the evaluation team, after being developed in consultation with the delivery team to ensure all key relevant aspects of LTSS were covered. They were designed to shed light on how the different participant groups perceived the delivery and impact of LTSS.

The case study visits took place during the second year of the programme to allow time to maximise awareness of the range of LTSS lessons and put teachers and students in a better position to take an overview of the approach. Two schools were selected at random from each of the three regions and all agreed to allow an evaluation visit. Further teacher feedback was gathered through on-line questionnaires with the URL sent to all schools (including controls) via the main contact person. This enabled teacher attitudes, approaches, and competencies to be compared between the two arms of the trial.

The teacher survey was administered towards the end of the two years (early summer term 2015) to compare responses as the programme became more familiar to teachers in the intervention schools. Topics covered by the questionnaire were finalised after discussion with the developers about their expectations of the programme, and included:

- teaching strategies used in science lessons (type and frequency); and
- confidence in teaching and learning approaches in science.

For LTSS schools only, they included:

- LTSS training attendance;
- number of LTSS lessons taught;
- frequency of using LTSS approaches/elements/practices;
- specific views on effectiveness of LTSS;
- challenging aspects of LTSS;
- suggested improvements to the LTSS programme;
- views on the impact of the LTSS programme on the students;
- whether the approach appears to be more or less suitable for any student groups (e.g. disadvantaged students); and
- any information about knock-on effects across the curriculum.

For control schools only, they included:

- views on the impact of science lessons on the students; and
- whether their approach to science teaching appears to be more or less suitable for any student groups (e.g. disadvantaged students).

LTSS schools were also sent a URL for students to complete an online survey, focusing on their experience of LTSS (enjoyability; lesson coverage and structure; teaching approaches) and also asking them to make some comparisons with their standard science lessons. The original intention, as outlined in the protocol, was to survey all students but when it because apparent how separate the LTSS intervention was from the regular science curriculum, it was decided to reduce the burden on the control schools by not including the students in the survey.

The surveys were analysed using both quantitative and qualitative techniques: frequency counts of Likert scales and thematic analysis of open questions. There was no explicit fidelity measure as part of this evaluation since the intervention did not incorporate a fixed programme of observations by the tutors or of self-assessment by the teachers. However, the responses to some of the survey questions could be used to draw broad inferences about the level of fidelity.

## Timeline

| Date | Activity |
|---|---|
| May – July 2013 | Recruitment of schools. |
| July 2013 | Randomisation |
| September 2013 – July 2015 | LTSS delivered to intervention schools (2 academic years) |
| March – May 2015 | Lesson observations |
| May 2015 | Parental opt-out consent for surveys and tests |
| May 2015 | Student surveys |
| May 2015 | Teacher surveys |
| June – July 2015 | Post-tests |
| September 2015 | Waitlist control schools receive LTSS |

## Costs

Cost information was collected from the project delivery team. Since this evaluation was conducted under pre-April 2013 EEF guidance, structured estimates were not collected from schools. However, from conversations with teachers and technicians during the evaluators' school visits, there was a suggestion that the burden (especially sourcing equipment by the technician) was somewhat more than the delivery team estimates.

# Impact evaluation

## Participants

Figure 2 shows the flow of participants through the trial. There were 53 schools at randomisation, 26 assigned to the intervention and 27 to the control. The exact number of students is not known since two schools (both in the LTSS condition) withdrew from the trial without providing UPN data for the study cohort. Of the 8,016 students in the 51 schools that did provide UPN data, some were lost due to four schools (one intervention and three control schools) failing to return any post-test data, and others from across the remaining schools not providing post-test data. Some of this would be due to student absence and moving schools. However, some schools returned incomplete data because they misunderstood the testing requirements. The project team were conducting their own testing and some teachers were confused about which students should undertake the independent evaluation, despite earlier guidance.

Of the five intervention schools that dropped out of the LTSS programme, four cited reasons related to staffing and the fifth had appointed a new headteacher who had different priorities. Three of the five schools agreed to continue as part of the ‘intention to treat’ sample but one failed to provide post-test data. Adding this school to the two who had not agreed to continue in the study made three schools that could not be included in the ITT analysis and two that were.

**Figure 2: Consort diagram**

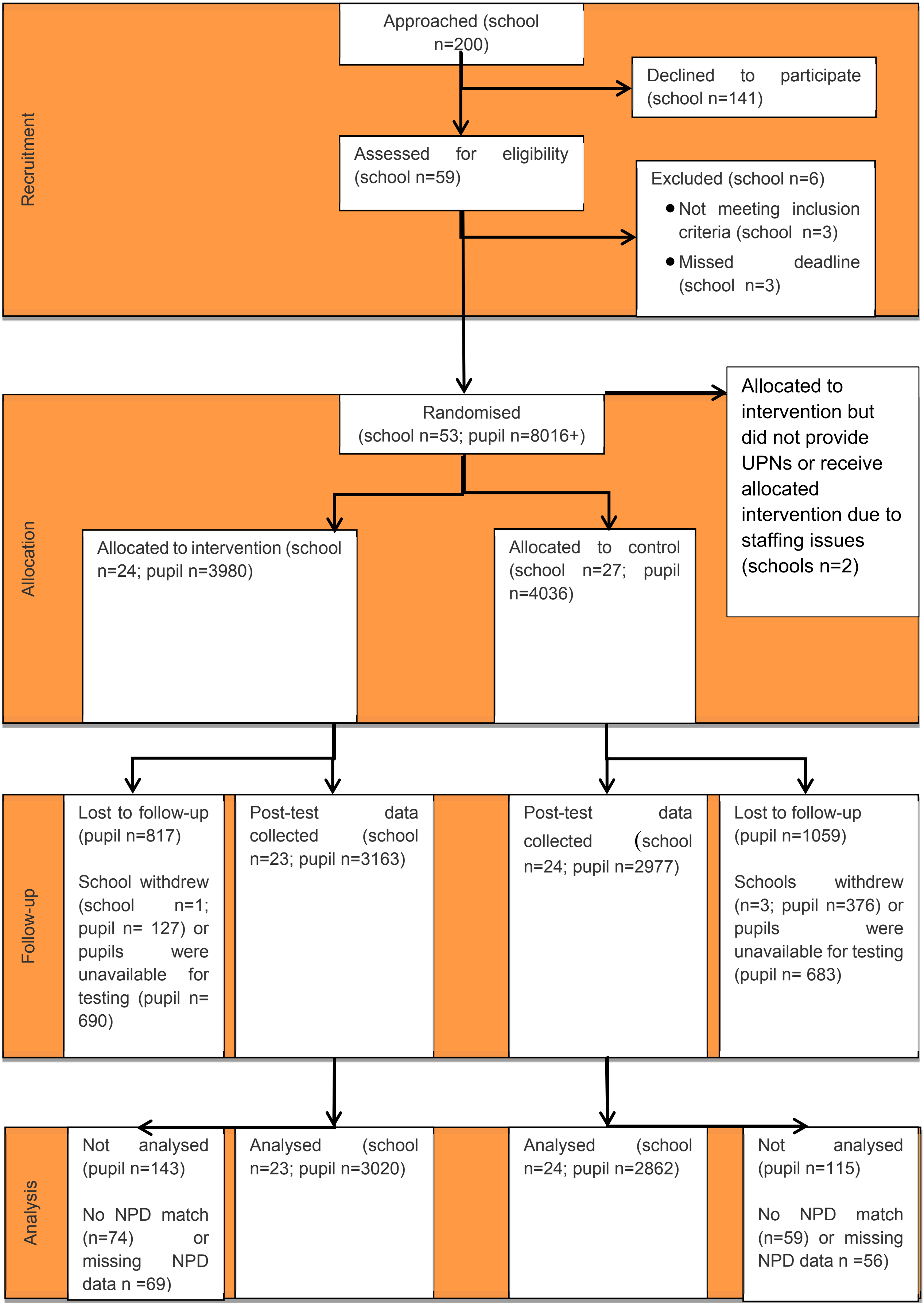

Recruitment
Approached (school n=200)
Declined to participate (school n=141)
Assessed for eligibility (school n=59)
Excluded (school n=6)
● Not meeting inclusion criteria (school n=3)
● Missed deadline (school n=3)
Allocation
Randomised (school n=53; pupil n=8016+)
Allocated to intervention but did not provide UPNs or receive allocated intervention due to staffing issues (schools n=2)
Allocated to intervention (school n=24; pupil n=3980)
Allocated to control (school n=27; pupil n=4036)
Follow-up
Lost to follow-up (pupil n=817)
School withdrew (school n=1; pupil n= 127) or pupils were unavailable for testing (pupil n= 690)
Post-test data collected (school n=23; pupil n=3163)
Post-test data collected (school n=24; pupil n=2977)
Lost to follow-up (pupil n=1059)
Schools withdrew (n=3; pupil n=376) or pupils were unavailable for testing (pupil n= 683)
Analysis
Not analysed (pupil n=143)
No NPD match (n=74) or missing NPD data n =69)
Analysed (school n=23; pupil n=3020)
Analysed (school n=24; pupil n=2862)
Not analysed (pupil n=115)
No NPD match (n=59) or missing NPD data n =56)

**Table 1: Minimum detectable effect size at different stages**

| Stage | N [schools] (n=intervention; n=control) | Correlation between pre-test (+other covariates) & post-test | ICC | Blocking/stratification or pair matching | Power | Alpha | Minimum detectable effect size (MDES) |
|---|---|---|---|---|---|---|---|
| **Protocol** | 53 (26; 27) | 0.75 | 0.12 | Pair matching | 80% | 0.05 | 0.20 |
| **Randomisation** | 53 (26; 27) | 0.75 | 0.12 | Pair matching | 80% | 0.05 | 0.19 |
| **Analysis (i.e. available pre- and post-test)** | 47 (23; 24) | 0.67 | 0.129 | Pair matching | 80% | 0.05 | 0.23 |

## Pupil characteristics

Before the analysis it was checked whether the available baseline characteristics were balanced across the two groups. For this, we estimated (generalised) linear mixed models predicting missingness of data and gender (generalised linear mixed model) and KS2 results (linear mixed model) which estimated random effects on school level and predicted these random effects with the treatment group. Treatment was not a significant predictor of missingness of data. The same was true of gender. The generalised linear mixed model for KS2 results did also not reveal any pre-study differences.

**Table 2: Baseline comparison**

| Variable | Intervention group | | Control group | |
|---|---|---|---|---|
| | N (missing) | Percentage | N (missing) | Percentage |
| **Type of school** | 26 (0) | (n/26)% | 27 (0) | (n/27)% |
| Academy converter | 8 | 30.77 | 9 | 33.33 |
| Academy sponsor led | 4 | 15.38 | 2 | 7.41 |
| Community school | 5 | 19.23 | 6 | 22.22 |
| Foundation school | 5 | 19.23 | 4 | 14.81 |
| Voluntary controlled school | 3 | 11.54 | 2 | 7.41 |
| Voluntary aided school | 1 | 3.87 | 4 | 14.81 |
| **Ofsted rating** | 26 (2) | % | 27 (5) | % |
| Outstanding | 4 | 17 | 3 | 14 |
| Good | 12 | 50 | 11 | 50 |
| Requires Improvement/Satisfactory | 7 | 29 | 7 | 32 |
| Inadequate | 1 | 4 | 1 | 5 |
| **School setting** | 26 (0) | % | 27 (0) | % |
| Urban conurbation | 8 | 31 | 11 | 41 |
| Urban city and town | 15 | 58 | 12 | 44 |
| Rural town and fringe | 3 | 12 | 4 | 15 |
| **Gender** | 26 (0) | % | 27 (0) | % |
| Mixed | 25 | 96 | 27 | 100 |
| Boys-only | 1 | 4 | 0 | 0 |
| Girls-only | 0 | 0 | 0 | 0 |
| **Number on roll** | 26 (0) | [Mean] | 27 (0) | [Mean] |
| Total school | 24049 | 925 | 23392 (0) | 867 |
| **Number on roll** | 26 (2) | | 27 (0) | |
| Year 8 | 3990 | 166 | 4043 | 150 |
| **Ever FSM** | 26 (0) | % | 27 (0) | % |
| Total Students | 5578.5 | 23 | 6166.5 | 26 |
| | 26 (2) | | 27 (0) | |
| Year 8 | 1078 | 27 | 1244 | 31 |
| **EAL** | | % | | % |
| Total Students | 1231 | 5.1 | 937 | 4.0 |

## Outcomes and analysis

This section gives a brief overview of the findings of the trial. Much fuller details of the analysis conducted and the outcomes can be found in the Technical Appendix to this report.

Analysis of the sub-sample of double-marked science tests showed that most of the variance in the test results was due to differences between students (78.8%) or between schools (20.8%) rather than differences in marking behaviour (0.3%). This can be seen as an acceptable result in terms of the reliability of the primary outcome.

Table 3 presents the findings for the primary outcome of the trial. The estimated mean values for both treatment groups are very similar and their confidence intervals overlap. The obtained effect sizes from analyses weighted for missingness as well as without the weighting can be classified as very small.

Repeating the analyses with just those schools that adhered to their random allocation to the LTSS or control groups (per protocol) showed again an effect size that can be classified as very small.

Overall, the difference between scores on the science attainment test (maximum score 90) for students in the LTSS and those in the control groups was very small (the intervention group scored just 0.13 points higher on average).

**Table 3: Primary analysis**

| | Expected means | | | | Effect size | | |
|---|---|---|---|---|---|---|---|
| | Intervention group | | Control group | | | | |
| Outcome | n (missing) | Mean (95% CI) | n (missing) | Mean (95% CI) | n in model (intervention; control) | Hedges g (95% CI) | p-value[2] |
| **Science test** | 3020 (128) | 46.29<br>(44.43, 48.16) | 2862 (285) | 46.42<br>(44.92, 47.93) | 5882<br>(3020; 2862) | -.01<br>(-.09, .04) | .75 |
| **Science test weighted** | 3020 (128) | 48.75<br>(46.88, 50.61) | 2862 (285) | 48.93<br>(47.37, 50.49) | 5882<br>(3020; 2862) | -.02<br>(-.10, .04) | .79 |
| **Science test, per protocol** | 2454 (694) | 46.60<br>(44.54, 48.66) | 2862 (285) | 46.46<br>(44.95, 47.96) | 5316<br>(2454; 2862) | .01<br>(-.08, .10) | .38 |
| **Science test, per protocol weighted** | 2454 (694) | 48.85<br>(46.80, 50.90) | 2862 (285) | 48.77<br>(47.21, 50.32) | 5316<br>(2454; 2862) | .01<br>(-.08, .10) | .38 |

Table 4 presents the results for the analyses of the secondary outcomes. No evidence was found in favour of a positive transfer effect (single-sided hypothesis testing). Contrary to expectations, the effect sizes were even negative which indicates the potential for negative transfer effects. However, two caveats must be taken into account when gauging the potential for negative transfer effects:

1. The study design was not robust enough to pick up effect sizes this small. For logistical reasons, half the schools administered maths and half English, rather than half the students in each school completing each test as in the original protocol. Some schools failed to return tests leading to sample sizes too small to draw robust conclusions (17 and 20 schools for maths and English respectively).
2. The analysis was appropriate only for identifying positive rather than negative effects (a one-sided hypothesis test). Consequently, only increases in scores were counted as transfer effects happening. So it is not statistically correct to cite this as conclusive evidence of negative impact.

[2] One-sided bootstrapped *p*-value

**Table 4: Secondary analysis**

| | Expected means | | | | Effect size | | |
|---|---|---|---|---|---|---|---|
| | Intervention group | | Control group | | | | |
| Outcome | n (missing) | Mean (95% CI) | n (missing) | Mean (95% CI) | n in model (intervention; control) | Hedges g (95% CI) | p-value[3] |
| **English test** | 964 (279) | 92.33 (87.41, 97.24) | 1433 (229) | 94.01 (90.35, 97.66) | 2397 (964; 1433) | -.15 [-.24, -.06] | .99 |
| **English test, weighted** | 964 (279) | 95.15 (90.34, 99.97) | 1433 (229) | 96.71 (92,99, 100.43) | 2397 (964; 1433) | -.13 [-.23, -.04] | .99 |
| **Maths test** | 816 (424) | 95.44 (92.35, 98.52) | 959 (89) | 96.41 (94.62, 98.20) | 1775 (816; 959) | -.11 [-.23, .01] | .96 |
| **Maths test, weighted** | 816 (424) | 97.96 (94.65, 101.27) | 959 (89) | 99.07 (97.16, 100.97) | 1775 (816; 959) | -.13 [-.25, -.02] | .98 |

In accordance with the protocol, the data was also analysed by a number of subgroups (Table 5). No impact was identified for any of these: low, medium and high attaining students (as defined by performance at KS2 maths and reading); boys; and those who had ever been eligible for FSM. There was a significant positive effect size within the group of female students only, indicating that female students who received LTSS did slightly better than the control group in the science assessments. The estimated effect size within this group was very small and did not reach the previously set minimally relevant effect size of Hedges *g* = .20.

Gender effects were further explored as indicated by EEF guidance as an interaction effect (a statistically more robust test to decide whether the effect of the treatment differs between female and male students). This analysis with the accompanying subgroup analyses (Tables 9 and 11, Appendix) revealed the following:

- The interaction between treatment and gender was statistically significant, indicating that the effect of LTSS differed between female and male students (Table 9, Appendix).
- The means for the Science Test Scores within female and male student samples did not differ significantly between those that received LTSS and the respective control groups (Table 11, Appendix), which also can be seen by the overlapping confidence intervals of these means within the two samples in Table 5.
- The mean Science Test Scores for male students receiving LTSS were slightly lower compared to the male control group, while the mean Science Test Scores for female LTSS students were slightly higher compared fo the female control group. These differences were too small to result in significant within group tests, which means that the differences within male and female sudents are likely to be too small to result in practically noticable differences (see also respective effect size estimates in Table 5 that are very small).

To summarise, considering (1) that the comparison of group means within subgroups did not show a difference within female or male students, as well as (2) the high number of exploratory subgroup tests performed, the significant effect size estimate could have occurred by chance and is statistically not robust. Further, the significant effect size found within the group of female student was smaller than hypothesised, i.e. potentially indicating a practically not relevant difference between female students receiving LTSS and the control group. Whether this effect within the group of female students holds up as well as whether there is a potential disadvantage of male students receiving LTSS as compared to female students receiving LTSS could nevertheless be a focus for future evaluations of LTSS.

[3] One-sided bootstrapped *p*-value

More detailed results on this as well as all other groups are reported in Appendix I.

**Table 5: Subgroup analysis**

| | Expected means | | | | Effect size | | |
|---|---|---|---|---|---|---|---|
| | Intervention group | | Control group | | | | |
| Outcome | n (missing) | Mean (95% CI) | n (missing) | Mean (95% CI) | n in model (intervention ; control) | Hedges g (95% CI) | p-value |
| **Science Test, FSM-only** | 783 (266) | 41.15 (39.12, 43.18) | 825 (293) | 41.44 (39.80, 43.08) | 1608 (783; 825) | -.03 (-.18, .06) | .84 |
| **Science test, female students** | 1441 (56) | 47.17 (45.37, 48.97) | 1400 (147) | 46.23 (44.87, 47.60) | 2841 (1441; 1400) | .10 (.03, .19) | .01 |
| **Science test, male students** | 1579 (72) | 45.89 (44.00, 47.76) | 1462 (138) | 46.43 (44.85, 48.01) | 3041 (1579; 1462) | -.05 (-.13, .03) | .89 |
| **Science test, KS2 stratum 1** | 967 (53) | 33.85 (31.86, 35.84) | 881 (121) | 33.15 (31.60, 34.69) | 1848 (967; 881) | .02 (-.04, .15) | .12 |
| **Science test, KS2 stratum 2** | 998 (41) | 45.79 (43.59, 47.99) | 991 (101) | 45.57 (43.72, 47.41) | 1989 (998; 991) | .05 (-.11, .10) | .53 |
| **Science test, KS2 stratum 3** | 1055 (34) | 58.11 (56.20, 60.02) | 990 (63) | 58.50 (56.66, 60.34) | 2045 (1055; 990) | -.04 (-.17, .05) | .84 |

## Cost

Since this was an evaluation conducted under the pre-April 2015 EEF guidance on cost evaluation, financial information was collected from the project delivery team. The main costs related to the training, with much lower amounts for lesson resources and equipment. Using figures provided by the delivery team, the cost of the intervention per school was calculated as shown in Table 6. All costs are start-up rather than running costs, and to keep the model simple, all costs have been front-loaded to the first year. In the second and third years, the cost is likely to be minimal (e.. replacing some resources) and has been counted as £0. The overall cost per school for the two-year programme was £4490. LTSS involved some burden of extra time, which the delivery team estimated at five hours for teachers (mainly planning) and two hours for a technician, per year. Since this evaluation was conducted under pre-April 2015 EEF guidance, structured estimates were not collected from schools, but there was a suggestion that the burden (especially on the technician to source equipment) was somewhat more than this.

**Table 6: Cumulative cost of LTSS (per school)**

| | Year 1 | Year 2 | Year 3 |
|---|---|---|---|
| **Total** | 4490 | 0 | 0 |
| **3 days' tutor time** | 3600 | 0 | 0 |
| **Tutor expenses** | 640 | 0 | 0 |
| **Resources** | 250 | 0 | 0 |

Since this is a two-year intervention, the cost per pupil has been calculated based on one cohort starting in Year 1 and finishing in Year 2, a second cohort in Year 2 to Year 3, and a further cohort half-counted in Year 3 (since their second year falls outside the three-year costing model). Each cohort consists of 150 students. The average cost is therefore £11.97 per pupil (4490/375). If all costs are paid in the first year, this equates to per pupil per year costs of £11.97 in the first year, falling to £5.99 per year in the second year and £3.99 in the third year (Table 7).

**Table 7: Cost per year over multiple years (pupils)**

| Number of years using programme | Cumulative cost per pupil | Average cost per pupil per year (cumulative cost per pupil/number of years) |
|---|---|---|
| **1 year** | £11.97 | £11.97 |
| **2 years** | £11.97 | £5.99 |
| **3 years** | £11.97 | £3.99 |

There were additional time requirements for the schools in participating in LTSS. As well as time for training (one full day in each of the two years, plus another three support days per year delivered in different patterns as negotiated with the school), there was also about five extra hours’ planning time per teacher and two hours for the technician per year.

The supply cover requirements would be one day per year for each participating teacher.

The delivery team estimated that schools would not be spending any additional money on equipment above that normally incurred by the science lessons LTSS was replacing. However, as explained above, we did not collect school information on this aspect.

Other than sourcing equipment they did not already have available, schools were not asked to contribute financially towards this project.

# Process evaluation

The process evaluation was designed to explore how LTSS was operationalised in schools, the reaction of teachers and students, and to make some comparisons between LTSS and standard science lessons. It is important to bear in mind when reading this section that the findings are drawn from a self-selecting sample of survey respondents and visits to 6 of the 21 intervention schools that completed the programme.

Judging from teacher self-report and student feedback as well as evaluators’ lesson observations, most teachers seemed to have adopted the essential composition of Let’s Think lessons, which featured group and paired discussions, time for students to think, and hands-on activity. Teachers showed a good grasp of the programme’s principles, but tended to be unsure whether it had improved students’ thinking skills.

More students preferred Let’s Think lessons to their standard science lessons and a majority rated them easy. There was a wide range of opinion in terms of enjoyment, with the main negative being that the lessons were boring and repetitive.

Although several teachers acknowledged that it took a while to embed pedagogical approaches such as facilitating small group work and non-directive questioning, there were instances where teachers felt such skills had improved and fed through to their teaching more broadly. Some felt the programme was too difficult for their low-attaining students, but this was not universal and the impact evaluation showed no evidence of a negative effect on these students. There were also some logistical challenges: fitting all the material into 50-minute lessons; finding time for planning; and sourcing some of the equipment.

The next two sections provide more detail about the evaluation.

## Implementation and fidelity check

A lesson observation schedule was designed in consultation with the project team to ensure that it covered all the elements considered essential to LTSS. Key conditions related to the role of the teacher (to model and facilitate with minimum intervention; and to encourage exploratory talk); the nature of student engagement (working in small groups; showing collaborative, supportive behaviour; and keeping on-task); and lesson structure (presenting the problem in a relevant, non-abstract way; having periods for review and reflection; devoting most of the time to activities and thinking; stretching students through cognitive challenge).

Four schools were identified where the treatment was not delivered to the whole year cohort as originally intended. In other words, some students received the intervention but others were closer to the ‘business as usual’ control condition. For instance, in one school just two teachers received LTSS training and delivered to their two classes because the senior manager identified it as *‘a good opportunity of having funded training’* for those two teachers, who were both newly qualified. Another of the four schools began with partial delivery but, because classes were reconstituted in Year 8, the whole cohort received LTSS lessons in the second year. It was not possible to get exact figures from all these schools, but indications were that only about a third of the students had received LTSS when delivery was partial. This would have had some confounding effect, although the ‘as is’ analysis (i.e. analysis by which treatment students actually received rather than which they were assigned to receive) suggests it did not change the outcome of the trial (see Appendix Table 8).

Around a third of students remembered doing 13 or more Let’s Think lessons but a similar proportion remembered 6 or fewer. The reasons for this could include low delivery rates by some teachers and schools; students missing lessons or joining the school part-way through Years 7/8; or simple inability to recall the lessons (for instance, a fifth said they were ‘not sure’ how many they had attended but ‘more than one’). Whatever the reason, this spread should be considered when interpreting the comments in the student survey.

Teachers were asked both about the number of different LTSS lessons they had taught from the cycle of 19, and about the number of different sessions they had taken (i.e. including lesson content they had taught more than once). Just over half the teachers had taught at least 13 different LTSS lessons, although a quarter had taught 6 or fewer. This did not necessarily mean that students would not receive all 19 lessons, since the survey was completed before the end of the year and also the teachers could have been sharing delivery with a colleague or could have joined the school partway through the two-year LTSS cycle. In total (including repeats), over two-thirds had taught more than 10 LTSS sessions (almost a third had taught more than 30sessions). In the teacher interviews, there was a consensus that it took several lessons to get accustomed to the approach, so it would seem most teachers answering the survey had reached this threshold (only four said they had taught five LTSS lessons or fewer in total). About a third of the LTSS teachers said they had completed all or most of the six main training days, and all or most of the five intersession tasks. About half had had all eight face-to-face meetings with the tutor. Enthusiasm for the training and tutor support in the six schools visited was high.

Most observed lessons showed good or excellent fidelity to the LTSS approach. Schools were delivering to the lesson plans, starting with the engaging, relevant ‘hook’ and then proceeding through a series of tasks and activities. Teachers showed varying degrees of ability at flexing these tasks to fit within the lesson and only two of the eight observed lessons explicitly incorporated a final, metacognitive section where students were encouraged to reflect on their thinking processes during the lesson, link ideas and bridge to related contexts. Teachers in schools with 50-minute lessons often struggled to get through the lesson plans. Virtually all the classes were divided into small groups of three or four as anticipated, although in one case the groups were larger (up to six) and in another the students worked mainly in pairs.

The LTSS model encouraged teachers to deliver each lesson twice, the first time to the year group above the trial cohort, so that there would be a beneficial effect of the practice. Only two in five of the teachers reported having done this very or quite often, and a quarter had never done so. Similarly, fewer than half had often planned lessons jointly with other teachers as recommended. The in-school interviews also found only limited evidence of practice lessons or joint planning. Nonetheless, this was more than double the prevalence of joint planning of ordinary science lessons in the control schools, so it appears that teachers spent more time planning LTSS lessons collaboratively.

Part of the LTSS process was to involve senior management in a monitoring role, including observing LTSS lessons. However, 43 teachers (out of 45 answering) said they had experienced this either never or not very often. In contrast, 23 out of 55 teachers in control schools said their lessons were very or quite often observed by senior managers. Sometimes there were disagreements within departments, for instance over whether LTSS should be used in mixed ability classes (as recommended) or the usual ability sets. In two of the six schools visited, there were also reports of teachers who had been involved with CASE previously failing to take LTSS adaptations on board or engage with it seriously, since they assumed this programme was identical. There was a sense that these issues were magnified when senior management was not fully involved.

As intended, the teaching approaches used in LTSS lessons were different from those used in the control schools for ordinary science lessons. This was particularly obvious for small group or paired discussions, used ‘very often’ by three-quarters of LTSS teachers and just a quarter of controls. They were also twice as likely to give students time ‘just to think’ (half saying ‘very often’) or to run ‘hands-on activities’ (three in five ‘very often’). To a lesser degree, whole-class discussions were also more common in LTSS classrooms. In contrast, whereas teaching scientific facts is reported as rarely happening in LTSS classrooms, half the control teachers do so ‘very often’. Likewise, only about one in ten LTSS teachers say they ‘very often’ explain ideas to students, compared with two-thirds of control teachers.

Student feedback on the teaching methods used in the classroom also suggested that the strategies and characteristics were in line with the LTSS approach, which involves at different stages teachers introducing ideas and facilitating discussion and feedback in whole-class sessions as well as in small groups. Small group work and full class discussions were both common features of the lessons (over 80% for each reporting that they happened quite or very often) with 69% saying they often worked with a partner. The vast majority (88%) said the teacher very or quite often talked to the whole class. Working in silence, as would be expected, was uncommon (75% said they never or hardly ever did so).

Teachers all showed a good grasp of the principles underlying the approach, although in some cases they admitted problems putting everything into practice (especially giving away control—*‘Not telling them the answers, and not having concrete facts that I need to teach them’*). The lesson observations confirmed this could be a struggle with weaker aspects including teachers modelling skills for the students (e.g. how to get the students sharing ideas or explaining clearly) and restricting their intervention to a minimum. Some reported a knock-on effect in their other teaching, for instance using more open-ended questions, class discussion, and being confident enough to follow a student’s line of thinking. They liked the relevance of the LTSS approach but found it difficult to assess whether it had led to improved student thinking. Some teachers thought the clarity of student written work had improved (e.g. test answers), but others had seen no sign of transfer.

A serious difficulty was finding the time for joint planning, inter-session tasks and the training sessions in a busy curriculum often without, they felt, full support from senior management.

Several schools had found it challenging to source so much non-standard equipment, and some technicians had found themselves being very creative or asking for help from other departments in the school (e.g. Design and Technology). There were complaints from three of the six schools visited that the lesson resources had to be corrected for typing errors and occasionally more major mistakes (with a suggestion that this was a bigger issue in the second year).

**Participant reactions**

Although several students understood that LTSS was an attempt to get them to think differently (*‘more about what you think instead of what you’re actually capable of doing’*) they found it difficult to articulate how (*‘they make you think in a fun way’*). Those students who were critical of the approach often complained that it was not about science, was too maths-focused, and did not teach them anything new (*‘we don't even need to use our brains as the stuff we do is like a science lesson in a nursery’*).

Almost a third of LTSS teachers thought low ability pupils had struggled with LTSS, which was much higher than the proportion for other ability groups and also higher than the proportion of control teachers saying low ability pupils struggled with ordinary science lessons. This should be considered alongside the impact evaluation data that showed no differential effect by attainment level. Although a majority of students (67%) rated the lessons easy, several survey participants wrote that they did not like LTSS lessons because, for example, *‘they are confusing and quite difficult’*. Some teachers thought these students found it hard to accept the concept of there being no right answer and were more comfortable with the structure of standard science lessons. However, it should be noted that some teachers felt the approach worked well with lower ability students (*‘my sort of really low ability set, it’s about increasing their confidence in just the discussion of being able to contribute and put forward idea’*”) and examples were seen in the lesson observations of adaptations being made for students with quite serious learning challenges.

Several teachers reported struggling with the non-directive questioning approach, although they have found the discussion elements particularly effective components of LTSS (*‘I find it challenging because I always have a set plan and I mind how well the lesson to go, you know, I know what the beginning is, I know what the end point is and that’s it, I want to get there in the end, so I find it challenging for myself because I just find myself saying I want you to be here, I want you to go there, and obviously that doesn’t* *work*'). Others had noticed an improvement in their questioning techniques and welcomed the change of role to teacher as mediator, standing back rather than leading (*'take the back seat and just explore'*).

Many of the students were enthusiastic about the LTSS lessons, and reported that they enjoyed being able to take part in practical activities as a change from writing and using textbooks. Several described the lessons as fun and they were also rated easier than usual science lessons. They embraced the lack of right and wrong answers and the opportunities for discussion. Asked in the survey whether they preferred Let's Think lessons or standard science lessons, the balance overall said they favoured Let's Think (53% versus 34%). However, the level of reported enjoyment ranged widely and at the extremes one in ten said they didn't enjoy them at all and another one in ten enjoyed them a lot. Two-thirds of students found the lessons either 'quite' or 'very' easy.

Having discussions and working in small groups are key aspects of LTSS. Most students liked these elements (73% and 79% respectively). They enjoyed the opportunity to share ideas, learn from others' explanations, have everyone participate, and work with, and get to know, new people. There were some complaints about being unable to choose your own group and other group members being disruptive. Some teachers echoed this by referring to the problem of managing behaviour and off-task talk in small groups.

The evaluation team observed that students' ability to work together and support each other varied quite widely. Group composition also varied: in some classes, the groups were self-selected rather than chosen by the teacher as recommended by the programme. There were examples of off-task behaviour and this was reflected in the survey with over half the students (59%) agreeing it was 'a good chance to chat to friends'. Similarly, interacting with friends was a prevalent reason for liking LTSS, although this was clearly not always off-task (*'you don't do much in lesson and we just talk to friends'*; *'You get to talk to your friends about it and that's good because if you get stuck they help you*').

By far the main negative expressed by students was that the LTSS lessons were boring and slow-paced. They found them repetitive, with only one question explored across a lesson (*'we seem to be doing the same things over and over again'*; *'how you always get the same type of sheets and you do basically the same thing every lesson apart from the actual thing you are learning about'*). Although some criticised the lack of practical activities or experiments, others identified the LTSS practicals as a positive.

The teacher survey revealed some interesting comparisons between LTSS practitioners and their counterparts in control schools. Although the sample sizes are small, there were indications that LTSS teachers were more confident in various aspects of science teaching during their ordinary science lessons (specifically not the LTSS sessions). This was especially true for assessing students' work and planning lessons. To a lesser extent, they felt more confident in explaining scientific ideas and questioning students effectively. No marked difference was seen for teaching scientific facts and science practicals, where the majority of all teachers were 'very confident'. The two sets of teachers also responded similarly to 'helping students discuss scientific ideas', with only around a third saying they were 'very confident' (most of the others were quite confident).

LTSS teachers were less likely to strongly agree that the students had made good progress in LTSS lessons than control teachers referencing ordinary science lessons. A considerably higher proportion of them strongly agreed that students actively collaborated in LTSS lessons and that the lessons improved their students' ability to reason than was true for the controls.

## Formative findings

There were three areas of adherence to the programme that were lower than might be expected, namely:

- the frequency of joint planning;

- teaching each lesson twice (the first time to a non-study cohort); and
- level of involvement of the senior manager.

Although fidelity to the approach was generally high in the eight lessons observed, the bridging section, where students are encouraged to link their learning to similar but broader contexts, was often absent. This may have been related to teachers' struggles to fit the content into the lesson slot available.

Various possible improvements were identified during the process evaluation:

- Make LTSS more accessible and engaging for lower ability students. Teachers often framed this as a request for more differentiation.
- Focus more on managing group dynamics and promoting productive discussion in groups in early stages of training the teachers. Although teachers and students were mainly positive about working in small groups, there were concerns related to disruptive behaviour and off-task talking.
- Re-order the sequence of lessons. The purpose was twofold: to put lessons which developed group work at the beginning, and to put more difficult content (such as the probability) at the end.
- More science-based lessons and less maths content were requested by teachers and students. The two lessons based on probability had caused particular difficulty.
- Provide editable files. This would have allowed teachers to correct errors or make the lesson more flexible. Science departments would also have preferred equipment that was more straightforward to source/create, or was available to purchase as an LTSS-specific pack.

# Conclusion

### Key conclusions

1. This evaluation provided no evidence that Let's Think Secondary Science improved the science attainment of students by the end of Year 8.
2. Students who received LTSS did worse than the control group on the English and maths assessments, but this result could have occurred by chance and we are not able to conclude that it was caused by the programme.
3. Many schools did not implement the programme as intended by the developer. In many schools, individual teachers delivered fewer than the full programme of 19 lessons and senior leaders were less engaged than prescribed by the programme.
4. Although most teachers were providing opportunities for students to work collaboratively, there was some evidence that more support to help teachers promote effective small group discussions would be welcomed.
5. Previous evaluations of CASE have suggested that it had longer-term impacts on academic attainment. Future research could examine whether LTSS also has a long-term impact by examining the GCSE results of the pupils involved in this evaluation.

The Let's Think Secondary Science programme, as implemented in this evaluation, did not improve students' attainment on a science test based on a KS3 SATs paper, nor on the secondary measures of English and Maths tests from GL Assessment, compared to control schools. Neither did FSM students, boys or girls, or students of different ability levels improve more than similar students in control schools.

The 19 one-hour lessons over two years, one day of initial training for teachers and six days of in-school support, were designed to model for teachers how to structure their lessons, i.e. to have students engage in collaborative learning, exploratory talk, to reflect on their learning, and to generalise it to other contexts.

## Limitations

Analysis for missing data shows that there was no bias due to attrition. Attrition rates were considerably higher for students (at 26%, assuming the two schools that did not supply UPNs had the working average of 150 students per year group) than schools (11%). This reduced the statistical power of the trial, as did the lower than expected correlation between the school means on the KS2 pre-test and the science post-test of 0.67 ($R^2 = 0.45$, compared with an anticipated value of 0.56). However, despite being clustered in three regions of England, the schools covered a reasonable geographic spread and a breadth of types of school and student, making them fairly representative of state schools across the country.

There were some issues with dosage and fidelity to the model that it was not possible to quantify. Based on teacher and student report, it was unclear whether all 19 LTSS lessons would be delivered in the two years, and in some schools there were classes who had not received any of the lessons.

In a departure from the original protocol, no cognitive reasoning measure was administered because none was identified as being suitably sensitive but not inherent to treatment. Although this was regrettable, it would not have changed the outcome of the trial as it was only a secondary outcome measure.

## Interpretation

Let's Think Secondary Science aims to improve students' performance by improving their thinking processes. The model is based on approaches that have been found to have high evidence of impact according to the EEF’s Teaching and Learning Toolkit: collaborative learning, metacognition, and self-regulation.

The primary outcome measure was an age-appropriate science test based on a KS3 SATs paper, and the secondary measures were English and Maths tests from GL Assessment. There was no evidence that schools that were assigned the LTSS approach improved their performance on any of these measures compared to control schools by more than chance.

The findings of this evaluation are very different from those of Adey and Shayer (1993) on the CASE intervention, which found an impact on student performance on cognitive development and GCSE science, maths and English. However, the CASE RCT was much smaller scale (24 classes in 9 schools) than the one reported here. There were other factors that could have led to this discrepancy. First, as mentioned above, CASE was a considerably more intense intervention than LTSS (30 lessons rather than 19 over two years), therefore the intervention students were exposed to a higher dosage. It would be optimistic to expect such infrequent sessions on generic cognitive objectives to have a robust impact on actual science learning over a two-year period (particularly as the LTSS students would lose 19 lessons from their standard science curriculum). Second, the CASE lessons were delivered to one class within each of the nine study schools, with the likelihood that this meant much more deliverer/developer contact may have led to higher fidelity of implementation than in the current implementation. Third, the CASE evaluation used a pre- and post-test of cognitive development that was inherent to treatment (i.e. used the same Piagetian basis as the intervention itself) so was more likely to find a result that favoured the intervention. Fourth, using the same test as a control measure at pre-test also provided stronger statistical control than was implemented in the current study, which used KS2 results as the baseline.

The process evaluation highlighted some concerns over lack of accessibility to, and engagement of, lower ability students. It also suggested that some teachers needed more support to encourage more productive small group discussion. Although group work was popular with most students, there were some complaints about disruptive behaviour. A sizeable minority of students found LTSS boring, mainly because of the repetitive nature of the lessons.

## Future research and publications

Some studies of CASE (e.g. Adey & Shayer, 1993; Shayer, 2000), although showing no immediate impact of the intervention on academic attainment, have reported delayed gains as evidenced by GCSE results for science, maths and English. It might therefore be advisable to conduct analysis of GCSE performance of the study cohort when they reach that stage. However, since this is a delayed treatment approach, control schools will have access to the LTSS training and resources so the control cohort may be subject to some diffusion/contamination.

# References

Adey, P. and Shayer, M. (1993) ‘An exploration of long-term far-transfer effects following an extended intervention program in the high school science curriculum’. *Cognition and Instruction*, 11: 1, 1–29.

Choi, B.S., Han, H.S., Kang, S.J., Lee, S.K., Kang, S.H., Park, J.Y. and Nam, J H. (2002) ‘Effects of a cognitive acceleration program on secondary school students’. *Journal of the Korean Association for Science Education*, 22: 4, 837–850.

Endler, L.C. and Bond, T.G. (2008) ‘Changing science outcomes: Cognitive acceleration in a US setting’. *Research in Science Education*, 38: 2, 149–166.

GL Assessment Progress in Maths Test Level 13. Available at www.gl-assessment.co.uk/products/progress-maths (accessed 4 January 2016).

GL Assessment Progress Test in English Level 13. Available at www.gl-assessment.co.uk/products/progress-test-english (accessed 4 January 2016).

Iqbal, H.M. and Shayer, M. (2000) ‘Accelerating the development of formal thinking in Pakistan secondary school students: Achievement effects and professional development issues’. *Journal of Research in Science Teaching*, 37: 3, 259–274.

Jones, M. and Gott, R. (1998) ‘Cognitive acceleration through science education: alternative perspectives’. *International Journal of Science Education*, 20: 7, 755–768.

KS3 Science Tier 3-6 Paper 2 (2009) Qualifications and Curriculum Authority. Available at www.stem.org.uk/elibrary/resource/30543/key-stage-three-science-tests-2009 (accessed 9 January 2016).

Maume, K. and Matthews, P. (2000) ‘A study of cognitive accelerated learning in science’. *Irish Educational Studies*, 19: 1, 95–106.

Mbano, N. (2003) ‘The effects of a cognitive acceleration intervention programme on the performance of secondary school pupils in Malawi’. *International Journal of Science Education*, 25: 1, 71–87.

McCormack, L., Finlayson, O.E.and McCloughlin, T.J. (2014) ‘The CASE programme implemented across the primary and secondary school transition in Ireland’. *International Journal of Science Education*, 36: 17, 2892–2917.

McDonald, R.P. (1999) *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum Associates.

Oliver, M., Venville, G. and Adey, P. (2012) ‘Effects of a cognitive acceleration programme in a low socioeconomic high school in regional Australia’. *International Journal of Science Education*, 34: 9, 1393–1410.

Shayer, M. (2000) *GCSE 1999: Added-Value from Schools adopting the CASE Intervention*. London: King’s College.

# Appendix A: Consent form

THE UNIVERSITY *of York*

**INSTITUTE FOR EFFECTIVE EDUCATION, THE UNIVERSITY OF YORK**

## Let's Think Secondary Science

### Information Sheet for Parents/Guardians

*Dear Parent/Guardian,*
We would like to request your permission for your child to take part in an educational research study. The following information explains why the research is being done and what it would involve for your child.

The Institute for Effective Education (IEE) is part of the University of York. It aims to find out what works in teaching and learning and why, and then use the evidence to improve education.

This study will assess the effectiveness of Let's Think Secondary Science, a series of science lessons taught alongside normal science lessons. They are designed to help pupils improve their thinking skills as well as their science achievement. We will conduct this study with Year 8 classes in 50 secondary schools in England. The headteacher of your child's school has agreed to participate in this study.

Schools were randomly assigned either to use Let's Think Secondary Science starting in September 2013 and ending in June 2015, or to be a comparison school, teaching science as usual.  Your school had an equal chance (50%) of being in either group. After June 2015, all participating schools, including the comparison schools, will be able to use Let's Think Secondary Science if they wish.

Pupils participating in the research study will be asked to complete some short assessments in the next few weeks. Your child's scores will be seen only by those who mark the assessments and by your child's teachers. Then pupils' names will be replaced with code numbers and no individual pupil's data will appear in any report about the research study. Pupils might also be asked for their opinion of Let's Think Secondary Science by filling in a survey or having a discussion with a researcher. Their answers will be completely confidential to the research team.

You may choose not to permit your child to take part in the research study, but (unless they are in the comparison group) they will have participated in the Let's Think Secondary Science lessons, as these have been part of science teaching throughout Year 7 and Year 8.

Please could you tell your child about the research study – that it aims to evaluate the effectiveness of Let's Think Secondary Science and will involve some short assessments - and explain that they have the right to withdraw from the assessments at any time.

If you do not want your child to participate, please complete and return the opt-out form below by [X pm on Y date]. A pupil's right to withdraw will be respected.

If you have a concern or question about your child's participation in this study, please contact Pam Hanley (e-mail: pam.hanley@york.ac.uk Tel:01904 328165) or Emma Marsden, the head of the Education Ethics Committee (email: emma.marsden@york.ac.uk) about the study.

Pam Hanley, Researcher, IEE, University of York

*************************************************************************
*********

**Parent/Guardian opt-out form**

If you **do not** permit your child to participate in the study, please complete this form and return it to your child's teacher by 3:00 pm on (one week after receipt).

I **do not** wish my child to take part in the research project. (If you do not want your child to take part, they will complete another piece of work set by their teacher when the other pupils are doing the tests)

**Pupil's name:** .......................................................................................
(Please print clearly)

**Form teacher's Name:** ....................................................................

**Parent’s/Guardian’s name:** ........................................................................
(Please print clearly)

**Parent’s/Guardian’s signature:** ..................................................................

**Date…**…………………………………………………………………………………………………

# Appendix B: School Memorandum of Understanding

## Let's Think Secondary Science Evaluation 2013-17

This randomised control evaluation aims to research the impact of Let's Think Secondary Science (LTSS). The programme has a particular emphasis on challenging pupils' thinking, developing student understanding in the key concepts that underlie scientific reasoning, and encouraging group learning. It will provide independent information for Headteachers and policy makers about the effectiveness of LTSS. If you agree to take part and accept the terms and conditions for receiving the resources and training then please sign a copy of this form and return it the address provided at the end of this letter.

All results will remain confidential, with no school being identified by name. Participating schools will be randomly assigned either as an "intervention" school (cohort 1) or as a "control school" (cohort 2), with a 50% chance of being assigned to either group. The evaluation will focus on students entering Y7 in September 2013.

Cohort 1 Schools will be provided with a training programme over 2 years and teaching resources free of charge. At the end of the first term of training, they will receive a contribution of £3000 for supply cover.

Cohort 2 schools will continue with their normal teaching during 2013-15 (while administering tests to the Years 7 and 8 students). They will be provided with the same training, resources and cover contribution (from September 2015).

As part of the EEF evaluation KS2 SATs results will be obtained from the National Pupil Database for both cohorts to provide the pre-test baseline. Students will be tested towards the end of Year 7 (one hour) and Year 8 (two hours). In a sub-sample of cohort 1 schools a researcher will visit to see how students are interacting with the programme. They will also talk to staff about their teaching and training experiences. A short online survey will also be administered to students and teachers in the cohort 1 schools.

<u>Requirements for schools</u>

I confirm that:

- Participating teachers have not received CASE (Cognitive Acceleration through Science Education) training in the previous 10 years.
- The school is not participating in another research project or programme evaluation (i.e. with science, cognitive development, literacy or numeracy outcomes) that would interfere with the LTSS evaluation.
- Teachers will attend training days and work with LTSS tutors throughout the project to ensure that they are implementing all aspects of the programme.
- For Cohort 1, each teacher will teach each lesson twice of which one class must be in Year 7 of the cohort.
- We allow the administration of Science Reasoning Tasks (SRT) and permit the publication of anonymised data collected.
- There will be access to Skype for video conferencing.

let's think cognitive acceleration

Education Endowment Foundation

Let's Think Secondary Science
Institute for
**Effective Education**

Requirements for schools: IEE independent evaluation

- Teachers will oversee the completion of tests for the relevant classes of pupils within the time frame and conditions specified by the IEE. These should be completed in exam conditions and invigilated by teachers who have not been involved in the LTSS research. There will be assessments at the end of Year 7 (1 hour) and at the end of Year 8 (2 hours). Some of these will be online.
- Teachers and pupils will complete an online questionnaire about aspects of the research in the second year of the evaluation.
- Teachers and schools will accommodate observational visits conducted by IEE staff. These will take place in a sub-sample of cohort 1 schools and involve lesson observations as well as discussions with teachers, students and a member of the senior management team.
- The school agrees to the IEE (or other evaluating bodies commissioned by the Education Endowment Foundation) obtaining data on the evaluation cohort's KS2 and GCSE results from the National Pupil Database, and will provide the UPNs to enable this to be achieved.
- If the school has to withdraw from the research for operational or other unavoidable reasons, wherever possible it will provide test data for the evaluation.
- Teachers will, at the earliest opportunity, notify the IEE if there are any support or operational issues preventing the effective use of the programme.
- Teachers will provide valid email addresses and telephone contact numbers to the researchers and tutors and agree to check communications regularly during the period of the research.
- In order to prevent data contamination while cohort 2 schools are receiving LTSS training, those teachers must not teach any LTSS lessons to Y9 in the academic year 2015/16.

Commitments of LTSS

- Training, teaching resources and ongoing support will be free of charge provided by LTSS to all schools. In addition £3000 will be provided as a contribution towards for teacher cover.

Commitments of the IEE

- All data provided by schools will be stored in accordance with the Data Protection Act (1998).
- All results will be anonymised so that no schools will be identifiable in the report or dissemination of results. Confidentiality will be maintained and no one outside the evaluation team will have access to the database.

***Headteacher agreement***

I agree for my school ________________________________________________ to take part in the LTSS evaluation and I accept the eligibility terms and conditions.

**Signature of Head Teacher:** ________________________________________

**Name of Head Teacher:**__________________________________________

**Date:** ___/___/______

**PLEASE RETURN TO Pam Hanley, IEE, Berrick Saul Building, University of York YO10 5DD;** pam.hanley@york.ac.uk**; or fax to 01904 328156**

# Appendix C: Teacher survey

## Let's Think Secondary Science evaluation<br>Teacher survey (intervention schools)

**1.** What is the name of your school? [mandatory, drop down list]

**2.** Do you teach science to Year 8 classes? [mandatory, one box]

□ yes – Let's Think and ordinary science lessons
□ yes – Let's Think lessons only
□ yes – ordinary science lessons only
□ no – terminate
□ other – please write in ______________________________

**3.** Thinking of ordinary science lessons rather than Let's Think lessons, how confident do you feel doing the following? [optional, one box per row]

| | | Very confident | Quite confident | Not very confident | Not at all confident |
|---|---|---|---|---|---|
| a) | Planning science lessons | □ | □ | □ | □ |
| b) | Assessing students' science work | □ | □ | □ | □ |
| c) | Teaching scientific facts | □ | □ | □ | □ |
| d) | Teaching science practicals | □ | □ | □ | □ |
| e) | Explaining scientific ideas | □ | □ | □ | □ |
| f) | Questioning students effectively | □ | □ | □ | □ |
| g) | Helping students discuss scientific ideas | □ | □ | □ | □ |

**4.** Approximately how many of the 19 different Let's Think lessons have you taught? (include only lessons you have taught to students in the research group ie now Year 8). [mandatory, one box]

□ 0
□ 1-3
□ 4-6
□ 7-12
□ 13+
□ not sure but more than one

**5.** And about how many Let's Think lessons have you taught IN TOTAL? (include lessons you have taught to students not in the research year ie now Year 8, and lessons you have taught more than once). [mandatory, one box]

□ 0 – STOP
□ 1-5

□ 6-10
□ 11-20
□ 21-30
□ 31+
□ not sure but more than one

**6.** How often have you done the following in the Let's Think project? [optional, one box per row]

| | | Very often | Quite often | Not very often | Never |
|---|---|---|---|---|---|
| a) | Practised the lesson with another class before using it with the research project students | □ | □ | □ | □ |
| b) | Planned lessons jointly with other teacher(s) | □ | □ | □ | □ |
| c) | Had lessons observed by a senior manager in the school | □ | □ | □ | □ |
| d) | Used techniques I learnt in Let's Think during ordinary science lessons | □ | □ | □ | □ |

**7.** How often do you do the following in Let's Think lessons? [optional, one box per row]

| | | Very often | Quite often | Not very often | Never |
|---|---|---|---|---|---|
| e) | Student hands-on activities | □ | □ | □ | □ |
| f) | Explain ideas to students | □ | □ | □ | □ |
| g) | Student discussion in whole class | □ | □ | □ | □ |
| h) | Student small group or paired discussion | □ | □ | □ | □ |
| i) | Give students time just to think | □ | □ | □ | □ |
| j) | Teach scientific facts | □ | □ | □ | □ |

**8.** Do you think Let's Think has been particularly beneficial for certain groups of pupils? Please tick which one(s) [optional]:

□ English as an Additional Language (EAL)
□ Special Educational Needs (SEN)
□ Disadvantaged pupils
□ Boys
□ Girls
□ Low ability
□ Middle ability
□ High ability
□ None in particular
□ Other (please write in): _______________

**9.** Do you find any particular groups of pupils have struggled with Let's Think? Please tick which one(s) [optional]:

□ English as an Additional Language (EAL)
□ Special Educational Needs (SEN)
□ Disadvantaged pupils
□ Boys
□ Girls
□ Low ability
□ Middle ability
□ High ability
□ None in particular
□ Other (please write in): _____________

**10.** Reflecting on your perception of your students' experience of Let's Think, how much would you agree or disagree with the following statements? [optional, one box per row]

| | | Agree strongly | Agree slightly | Disagree slightly | Disagree strongly |
|---|---|---|---|---|---|
| a) | My students have engaged positively with Let's Think lessons | □ | □ | □ | □ |
| b) | My students have made good progress in Let's Think lessons | □ | □ | □ | □ |
| c) | My students are confident in Let's Think lessons | □ | □ | □ | □ |
| d) | Many of my students have found it too challenging | □ | □ | □ | □ |
| e) | It has improved my students' ability to reason | □ | □ | □ | □ |
| f) | My students actively collaborate with their peers | □ | □ | □ | □ |
| g) | Let's Think has helped my students understand their thinking processes | □ | □ | □ | □ |

**11.** There were three main components to the Let's Think training. How much of each did you manage to attend? [optional, one box per row]

| | All/most | Some | None |
|---|---|---|---|
| 6 main training days | □ | □ | □ |
| 8 face-to-face meetings with Let's Think tutor | □ | □ | □ |
| 5 intersession tasks | □ | □ | □ |

**12.** Please tell us any aspect(s) of Let's Think you have found particularly effective, and why.

**13.** Have you struggled with any aspect(s) of Let's Think?

**14.** Are there any improvements you would suggest to the programme?

*Many thanks for your help. Now please press submit.*

# Appendix D: Student survey

**Let’s Think Secondary Science evaluation**
**Student survey (intervention schools)**

1. What is the name of your school? [mandatory, drop down list]
2. Are you a girl/boy? [mandatory, one box]
3. Your school has been running a new kind of lesson in the last two years called Let’s Think Secondary Science. About how many of these lessons do you remember doing? [mandatory, one box]

   □ 0 – stop
   □ 1-3
   □ 4-6
   □ 7-12
   □ 13+
   □ not sure but more than one

4.
   a. How does Let’s Think Secondary Science compare with your other Science lessons? [Let’s Think Secondary Science lessons are … a lot more enjoyable, a little more enjoyable, a little less enjoyable, a lot less enjoyable, not sure] [optional, one box]
   b. Why do you say that?

5. How much do you enjoy your Let’s Think Secondary Science lessons? [a lot, quite a lot, not much, not at all] [optional, one box]
6. How difficult do you find what you do in Let’s Think Secondary Science lessons? [very difficult, quite difficult, quite easy, very easy] [optional, one box]
7. Please tick to show how much you agree or disagree with each statement about your Let’s Think Secondary Science lessons [agree a lot, agree, disagree, disagree a lot]: [optional, one box per line]
   a. we cover too much in each lesson
   b. there are no right or wrong answers
   c. I carry on thinking about the ideas in the lesson after it has finished
   d. the speed of the lessons is too slow
   e. I get lots of chances to share my ideas
   f. I feel confident about what I’m doing in these lessons
8. How often do you do each of the following in your Let’s Think Secondary Science lessons? [very often, quite often, hardly ever, never] [optional, one box per line]
   a. work with a partner
   b. work in small groups
   c. have full class discussions
   d. listen to the teacher
   e. work in silence

9. How much do you like each of these activities? [a lot, quite a lot, not much, not at all] [optional, one box per line]
    a. working with a partner
    b. working in small groups
    c. having discussions
    d. listening to the teacher
    e. working in silence
10. What do you like most about your Let’s Think Secondary Science lessons?
11. And what do you like least about them?

*Thank you for completing this survey.*

# Appendix E: LTSS lesson observation schedule

School: ________________________________ Teacher: ____________________________________________

Class: ____________________________________ Observer: ________________ Date: _______________

| | | | | | | |
|---|---|---|---|---|---|---|
| | 0 (never) to 3 (always/whenever appropriate) | | | | | Comments/notes |
| 1.Lesson starts with engaging hook | Yes | | | No | | |
| 2.The problem is presented in a relevant way (not abstract) | N/A | 0 | 1 | 2 | 3 | |
| 3. There are periods of review/generalisation where ideas are shared | N/A | 0 | 1 | 2 | 3 | |
| 4.Paced so majority of time is activities and thinking | N/A | 0 | 1 | 2 | 3 | |
| 5.Most students are stretched by the cognitive challenge | N/A | 0 | 1 | 2 | 3 | |
| 6.Students link ideas to current and future experiences and learning | N/A | 0 | 1 | 2 | 3 | |
| 7. Minimum teacher intervention unless needed to re-engage students or model group discussion | N/A | 0 | 1 | 2 | 3 | |
| 8. The classroom climate supports challenge and risk-taking. Exploratory talk is encouraged. | N/A | 0 | 1 | 2 | 3 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 9.Teacher models skills eg how to get people to share ideas, explain clearly, if necessary | N/A | 0 | 1 | 2 | 3 | |
| 10. Teachers and students develop and share vocabulary to facilitate communication of ideas | N/A | 0 | 1 | 2 | 3 | |
| 11.Pupils work in teams/groups | N/A | 0 | 1 | 2 | 3 | |
| 12.Pupils display collaborative behaviour during group work (e.g. active listening, everyone participates, explaining ideas to each other) | N/A | 0 | 1 | 2 | 3 | |
| 13.Pupils support each other | N/A | 0 | 1 | 2 | 3 | |
| 14.Pupils remain involved in the problem with few off-task behaviours | N/A | 0 | 1 | 2 | 3 | |
| 15.Pupils spend time reflecting and may leave lesson still puzzling | N/A | 0 | 1 | 2 | 3 | |

Overall Implementation Rating: 0 1 2 3

(0=LTSS in name only, no fidelity to the programme materials or teaching/ learning approach; 1=LTSS materials are in place, but the programme is not consistently/universally followed with fidelity. Teaching/learning approach is patchy; 2=LTSS materials and routines are followed with fidelity. Teaching/learning approach is not consistent; 3=LTSS materials and routines are followed with fidelity, and teaching/learning approach is embedded in the lessons)

v2 8/3/15

# Appendix F: Teacher interview protocol

**LTSS school visit: Interview with LTSS teachers**

1. Were you involved in the decision to take part in the trial of Let’s Think? IF YES: Why were you initially attracted to the Let’s Think programme? How did you think it might benefit your students?

2. How has the Let’s Think programme delivered compared with your expectations? (Probe benefits mentioned in previous question) IF NO AT Q1 ask instead: What benefits do you think the Let’s Think programme has delivered to your students?

3. Have there been any particular challenges as regards introducing Let’s Think in your school?

4. What do you think of the training and support package that has been offered by the Let’s Think Forum?

5. How many staff have been involved in delivering Let’s Think? (Establish what proportion of Y8 cohort)

6. Do you teach Let’s Think yourself?

7. Are there any aspects of Let’s Think that you think are particularly effective?

8. And are there any aspects that you think need changing or improving?

9. (IF NOT MENTIONED): What sort of peer support have you had? PROMPT: Have you been involved in peer-to-peer coaching?

10. (IF NOT MENTIONED): What sort of support have you had from the senior leadership team?

11. How have the students reacted to Let’s Think?

12. Have you any other comments?

### LTSS school visit: Interview with SLT teachers

1. Were you involved in the decision to take part in the trial of Let’s Think? IF YES: Why were you initially attracted to the Let’s Think programme? How did you think it might benefit your students?

2. How has the Let’s Think programme delivered compared with your expectations? (Probe benefits mentioned in previous question)

3. Have there been any particular challenges as regards introducing Let’s Think in your school?

4. What do you think of the training and support package that has been offered by the Let’s Think Forum?

5. Can you describe your involvement in Let’s Think? PROMPT IF NECESSARY: Do you teach it yourself? Do you get involved with teachers who do? What’s your role with them?

6. Are there any aspects of Let’s Think that you think are particularly effective?

7. And are there any aspects that you think need changing or improving?

8. Have you any other comments?

9. If not mentioned, ask:

   a. if there is an in-school learning group for Let’s Think

   b. how the school is monitoring the progress of Let’s Think

10. How have the students reacted to Let’s Think?

# Appendix G: Student focus group protocol

**LTSS school visit: Student focus group topic guide**

1. [If **after** lesson obs] Tell me about the science lesson you just did?

   **Probes:** What did you learn?
   What bits did you like best?
   Were there any things you found particularly difficult?

2. [If not yet mentioned]: I noticed you did quite a bit of work in groups.

   **Probes:** What do you like about working in groups?
   What don't you like so much?
   Does working in groups make it easier or harder to learn?
   Why?

3. I understand this was one of a series of lessons you have called "Let's Think Secondary Science". Is that right? (Might need to probe if they call it something else) IF YES: How do these lessons compare with your other science lessons?
   **Probes:** Which sort of lessons do you prefer? Why?

4. [If **before** lesson obs] I understand this was one of a series of lessons you have called "Let's Think Secondary Science". Is that right? (Might need to probe if they call it something else) IF YES: How do these lessons compare with your other science lessons?
   **Probes:** What do you learn?
   What bits do you like best?
   Are there any things you find particularly difficult?
   Which sort of lessons do you prefer? Why?

5. [If not yet mentioned]: I understand you do quite a bit of work in groups during LTSS lessons. IF YES:

   **Probes:** What do you like about working in groups?
   What don't you like so much?
   Does working in groups make it easier or harder to learn?
   Why?

# Appendix H: Statistical plan

## Statistical Analysis Plan

## Let's Think Secondary Science

Jan R. Böhnke, Tim J. Croudace, Pam Hanley, Louise Elliott

Draft 1.2, 25 April 2016

### I) Data Cleaning Procedures and Sampling Frame Definition

The data of the Science Test (based on KS3 Science Tier 3-6 Paper 2) will be entered into an ACCESS database. The data will be input twice: Input 1 requires the entry of the score for each individual question plus the total; Input 2 requires the entry of the total score. The scores will be checked to ensure they are within range and the total scores from input 1 and input 2 will be cross-checked. Discrepancies will be resolved by going back to the original data. To assess the reliability of the marking of the Science Test, 10% of the test papers (chosen at random) will be double marked and the intra-class correlation within this sample will be assessed as a measure of accuracy.

The data for the PIM and PTE (GL Assessment) are collected online for $N = 25$ schools and will be entered at GL for the remaining $N = 12$ schools. The paper versions will be scored manually by two markers, and in the case of any disputes a third marker will be used. The IEE LTSS team will download a csv file of the online data and GL will send through a csv file of the manually entered records. Both the online and paper PIM have a maximum mark of 50; however, the online PTE maximum is 67 and paper PTE maximum is 66. GL has recommended that the standard age scores[4] are used for these tests to ensure the online and paper versions are comparable.

Key stage 2 results will be obtained from the National Pupil Database (NPD), along with the pupil’s gender and FSM status. We will use EverFSM as specified by the EEF.

The IEE LTSS team will collate all the data to create a single SPSS dataset with one row per pupil that contains:

(a) the individual question scores and the total raw score for the science test

(b) question scores, total raw score and standardised score for the PTE or PIM as appropriate;

(c) the key stage results from the NPD including all relevant demographic information needed for sub-group analyses (see below II-1: free school meals, gender, and attainment based on KS2 at pre-test) as well as the following variables that are needed to test the effect of the intervention on achievement:

- an identifier variable for the schools;
- an identifier variable for the pupil;
- a variable coding whether the pupil attended a target school (where the intervention should have been delivered) or not (ITT intervention code);
- a variable coding whether the intervention was actually delivered or not (per protocol code).

Also, the total scores of the double marking of the Science Test should be provided for those records that were randomly sampled for this procedure as well as a code for first and second marker. This information is needed for the reliability analysis of the Science Test marks.

[4] These GL tests are standardised so that the average standard age score for any age group is always 100. The standard deviation is also set to plus or minus 15 points, so that for any age group about two-thirds of the students in the national sample will have a standardised score of between 85 and 115, approximately 95 per cent score between 70 and 130, and over 99 per cent score between 60 and 140. Raw scores are converted to standard age scores that allow you to compare the level of cognitive development of an individual with the levels of other students in the same age group.

After pseudonymisation this SPSS dataset will be made available through a shared network drive within the intranet of the University of York, with limited access managed by the IEE LTSS team. All analyses conducted on this dataset will be documented with reports and syntaxes on that shared drive. The final dataset, along with any derived variables, and the analysis syntax will be deposited with the EEF after submission and approval of the final report.

The analysis population is defined as follows:

- Our intention-to-treat sample consists of all the pupils whose details were provided by the participating schools (collected during the first year of the project).
- Any pupil from the NPD that provided at least one of the post-tests will be treated as "available for analysis". Students who only complete Part 1 of the two-part PTE will be excluded from PTE analysis because GL is unable to provide a standardised score (this includes one school where all 60 of the students doing the test only completed Part 1). Any pupil that could **not** be matched with the NPD and therefore not having KS2 results (pre-test) will not be used in the analyses (outside the sampling frame).
- Any pupil that was found in the NPD database but has not reported any results on any of the tests will be treated as a "drop-out" (panel attrition). Descriptive statistics will be used to compare intention-to-treat, per protocol and attrition cases. Multilevel models will be used to test whether school-specific effects on attrition are observed.

### II) Statistical Analysis Plan

The analysis for this project is divided in three parts. The *first part* is the test of the core hypothesis: Did the Let's Think Secondary Science program improve Science Achievement as measured by the Science Test. The *second part* is the exploratory analysis whether positive transfer effects can be observed as well in the PTE & PIM. The *third part* is the exploratory analysis of whether the three achievement tests assess a common latent dimension and how much of the variance in achievement is actually due to the different domains these tests assess.

Prior to all analyses outlined below the variables will be tested for the degree of normality realised and all analyses will be based on bootstrapped analysis to accommodate potential violations of normality. Also, the relationships between pre-test control variables and post-test outcomes will be checked for violations of linearity and appropriate transformations will be used if necessary. In this case, both the results from transformed as well as untransformed analyses will be reported.

*II-1) Analysis of primary outcome*

The Science Test was used to assess science achievement across all schools in the analysis sample and is the primary outcome measure to evaluate the effectiveness of the intervention. For this analysis the raw total scores will be used. The science test consisted of 15 questions (a mix of multiple choice, single word and short answer) split into between 2 and 7 sub-questions. Marks were awarded for each sub-question and the total calculated. Questions had total scores of between 4 and 8 marks. The sub-question total and overall total were entered into the database. Maximum total score was 90 (Sc1=33; Sc2=20; Sc3=16; Sc4=21: these sub-categories were not used for analysis). The main hypothesis of this project can therefore be formulated as a single directional hypothesis to decide about the effect of the intervention as well as its size:

*H1: Did the LTSS program improve science achievement as measured?*

This hypothesis will be tested using a hierarchical linear model, treating the children as nested within their respective schools. Control variables that will be used in this analysis are gender and the combined reading, writing and mathematics scores from the key stage assessment, controlling for potential baseline differences.

The variables will be centred at grand mean level. This way the intercept estimated for every school is the predicted science achievement level at the overall sample's mean achievement level. In the multilevel model a fixed effect will then be tested to evaluate whether the intercepts (representing mean science achievement at a given school) are higher for those schools that were randomised to the intervention than for those, that were randomised to the control condition. A conventional p-level of $p = .05$ will be applied and the bootstrapped p-value of the intervention coefficient will be used as a decision criterion (software used: Mplus). If this fixed effect is positive and significant, the LTSS program had a positive effect on Science Achievement.

*Sensitivity analyses:*

Sensitivity analyses are used to explore the robustness of the effect under several conditions. These cannot be used to judge the effectiveness of the intervention, they only inform about the limits of generalisability.

- The first sensitivity analysis will also use interaction effects between the treatment variable and the two key stage tests to explore whether the intervention has differential effects based on pre-intervention attainment levels.
- The second sensitivity analysis is the test of the *actual* administration of the intervention. Instead of using the intervention randomisation variable as in the primary endpoint analysis outlined above, information from the schools will be used regarding which pupils actually received the intervention. This will generate an upper-bound estimate of the intervention effect, which will be upwardly biased due to self-selection of the schools into the two groups.

*Subgroup analyses*

As with the sensitivity analyses, these subgroup analyses are only conducted to explore the limits of generalisability of the results and to identify potential inequalities in the effects. These analyses cannot be used to evaluate the LTSS program as effective or not. Seven subgroup analyses will be conducted:

- Comparable to the first of the sensitivity analyses the sample will be split based on key stage assessment scores into high vs. medium vs low attainment groups (terciles) and the same analysis as outlined above will be run within each of these achievement strata.
- The analysis will be run only in the subgroup of those schools that implemented the treatment they were randomised to (per-protocol analysis).
- The analysis will be run only within the subgroup of pupils who have ever been eligible for free school meals.
- And the sample will be split by gender and the analysis will be re-run separately for boys and girls.

*II-2) Analysis of secondary outcomes*

In addition to the analysis of the primary outcome, the PIM and PTE were collected in 17 and 20 schools respectively.

Table 1: Number of schools per secondary outcome and collection mode.

| Frequency of schools | PIM | PTE |
|---|---|---|
| Online | 12 | 13 |
| Paper | 5 | 7 |
| Not done | 8 | 6 |

To investigate whether differential and/or positive transfer effects can be detected across the different domains that these tests assess, the hierarchical linear model from step II-1 will be extended to include PIM and PTE scores as correlated dependent outcomes (since the PIM and PTE scores are likely to be highly correlated; e-mail communication with GL, Cres Fernandes, 28.06.2013). The same control variables will be used, but in addition it will be controlled for whether the PIM and PTE were administered online or as paper-pencil versions. Again, a positive effect of the LTSS intervention on all three outcomes will be seen as further corroboration that the LTSS intervention had a positive effect on Science attainment and in addition positive transfer effects on Mathematical and English Attainment could be observed.

Since these tests were not administered in all schools because of the assessment design, several changes to the analytic strategy have to be made. The missing observations on the not administered tests can be treated as missing by design, especially since schools have undergone a stratified randomisation procedure. But since the school (i.e. clustering variable) is correlated with whether a certain test was administered or not, multilevel analyses are conducted on the assumption that the structural parameters for the correlation of Maths and English attainment are the same within as well between schools.

### *II-3) Optimising psychometric strategies*

To test whether the three instruments used in this study actually assess different latent constructs, they will be modelled with a bifactor strategy with one general factor across all instruments and three factors for the specific tests. The general factor in this analysis represents a component of general achievement, common to all three ability domains (English, Maths, Science). The specific factors in turn assess any variability in outcomes that can be attributed to specific aspects and abilities of the three domains. This analysis is in a first step performed purely to explore the psychometric properties of the set of three instruments and to answer questions like: How many dimensions do these three tests actually assess and do they (and especially the Science Test) assess variance that is specific and useful beyond a general achievement dimension.

In a second step, a structural equation model will be used with the four latent scores (general achievement; specific English, Maths and Science achievement) as dependent variables to test differential intervention effects. As in the analysis for the primary endpoint, the four latent scores will be regressed upon the intention-to-treat variable of the intervention, while controlling for gender and attainment prior to the intervention (key stage test scores). The significance of the regression weights predicting the four latent scores based on the intervention variable will show whether the LTSS program has an effect specifically on Science Achievement (the latent scores on the specific factor for Science Achievement will differ significantly) or whether the LTSS intervention had an effect rather on general achievement (i.e. the general factor).

Since these tests were not administered in all schools because of the assessment design, several changes to the analytic strategy compared to the primary outcome have to be made. The missing observations on the not administered tests can be treated as missing by design, especially since schools have undergone a stratified randomisation procedure. But since the school (i.e. clustering variable) is correlated with the fact whether a certain test was administered or not, multilevel analyses are not possible. Therefore, the sample will be analysed as a single sample and no subgroup analyses will be performed on this tertiary outcome. The design effect introduced by this change of strategy will be quantified by reporting the intra-class correlations of the measures, ie, the amount of variance that is attributable to differences between schools.

# Appendix I: Detailed data analysis

### DETAILED DATA ANALYSIS

### DESCRIPTIVE ANALYSIS OF MISSINGNESS IN SPSS

*1) Preliminary Data Handling within SPSS*

A copy of the original SPSS file was used to calculate new variables that will be used as indicators in this analysis.

In a first step the number of valid cases and the number of relevant drop-outs need to be determined. The transferred data file contains $N_{\text{student}}$ = 8016 students in $N_{\text{school}}$ = 51 schools. For *n* = 102 no gender was documented and for *n* = 139 no documentation regarding free school meals was available. For n = 125 no valid KS2 results were available. Overall, for *n* = 142 (1.77%) no or only incomplete NPD data were available. Since this were only a small number of cases and they were most likely to have joined the schools later / during the intervention, they were not considered to be part of the sampling frame (see also analysis plan). Further, since these are all independent variables with relatively little overlap, no imputation or other algorithms are available that would help to correct for these. Schools 7 and 45 are the only schools for which a very high percentage of occurrences was reported.

Table 2: Percentage of missing NPD information by school

| School | Complete | Missin<br>g | Total | Percent<br>Missing<br>NPD |
|---|---|---|---|---|
| 1 | 138 | 4 | 142 | 2.90 |
| 2 | 73 | 2 | 75 | 2.74 |
| 3 | 238 | 0 | 238 | 0.00 |
| 4 | 170 | 0 | 170 | 0.00 |
| 5 | 160 | 3 | 163 | 1.88 |
| 6 | 73 | 1 | 74 | 1.37 |
| 7 | 65 | 26 | 91 | 40.00 |
| 8 | 108 | 4 | 112 | 3.70 |
| 9 | 124 | 2 | 126 | 1.61 |
| 10 | 191 | 6 | 197 | 3.14 |
| 11 | 97 | 0 | 97 | 0.00 |
| 12 | 124 | 4 | 128 | 3.23 |
| 13 | 117 | 0 | 117 | 0.00 |
| 14 | 137 | 2 | 139 | 1.46 |
| 15 | 243 | 10 | 253 | 4.12 |
| 16 | 173 | 0 | 173 | 0.00 |
| 17 | 270 | 4 | 274 | 1.48 |
| 18 | 269 | 7 | 276 | 2.60 |

| | | | | | |
|---|---|---|---|---|---|
| | 19 | 229 | 8 | 237 | 3.49 |
| | 20 | 115 | 2 | 117 | 1.74 |
| | 21 | 203 | 3 | 206 | 1.48 |
| | 22 | 58 | 0 | 58 | 0.00 |
| | 23 | 85 | 5 | 90 | 5.88 |
| | 24 | 100 | 0 | 100 | 0.00 |
| | 25 | 68 | 1 | 69 | 1.47 |
| | 27 | 91 | 0 | 91 | 0.00 |
| | 28 | 112 | 0 | 112 | 0.00 |
| | 29 | 127 | 0 | 127 | 0.00 |
| | 30 | 223 | 0 | 223 | 0.00 |
| | 31 | 204 | 1 | 205 | 0.49 |
| | 32 | 154 | 4 | 158 | 2.60 |
| | 33 | 197 | 4 | 201 | 2.03 |
| | 35 | 169 | 3 | 172 | 1.78 |
| | 36 | 54 | 0 | 54 | 0.00 |
| | 37 | 46 | 3 | 49 | 6.52 |
| | 38 | 190 | 1 | 191 | 0.53 |
| | 39 | 120 | 2 | 122 | 1.67 |
| | 40 | 177 | 0 | 177 | 0.00 |
| | 41 | 225 | 0 | 225 | 0.00 |
| | 42 | 197 | 2 | 199 | 1.02 |
| | 43 | 244 | 5 | 249 | 2.05 |
| | 44 | 93 | 1 | 94 | 1.08 |
| | 45 | 92 | 11 | 103 | 11.96 |
| | 46 | 268 | 3 | 271 | 1.12 |
| | 47 | 202 | 4 | 206 | 1.98 |
| | 48 | 214 | 0 | 214 | 0.00 |
| | 49 | 78 | 1 | 79 | 1.28 |
| | 50 | 164 | 1 | 165 | 0.61 |
| | 51 | 230 | 0 | 230 | 0.00 |
| | 52 | 208 | 1 | 209 | 0.48 |
| | 53 | 167 | 1 | 168 | 0.60 |
| Total | | 7874 | 142 | 8016 | 1.80 |

The variable *validsci* codes the number of valid responses per student and the variable *nosci* is an indicator variable that codes “1” if the student in question did not provide any responses to the science test and “0” otherwise. The distribution of this variable identifies the four schools that did not take part in the second assessment (IDs 5, 20, 29, 44) and it also shows the high variability of students that were enrolled at this particular school (overall N = 8016; last column) and the students that were actually assessed (N = 5882; second column). These students are the available / observed members for the intention to treat analysis. The additional members that provided NPD data will be used to generate the weight for providing no post-test (sensitivity analysis).

Table 3: Missing post-test (Science Test) data across schools for students with NPD data

| SchoolID | Science Test | No Science Test | Total |
|---|---|---|---|
| 1 | 127 | 11 | 138 |
| 2 | 66 | 7 | 73 |
| 3 | 194 | 44 | 238 |
| 4 | 136 | 34 | 170 |
| 5 | 0 | 160 | 160 |
| 6 | 63 | 10 | 73 |
| 7 | 62 | 3 | 65 |
| 8 | 94 | 14 | 108 |
| 9 | 85 | 39 | 124 |
| 10 | 167 | 24 | 191 |
| 11 | 86 | 11 | 97 |
| 12 | 116 | 8 | 124 |
| 13 | 99 | 18 | 117 |
| 14 | 115 | 22 | 137 |
| 15 | 215 | 28 | 243 |
| 16 | 154 | 19 | 173 |
| 17 | 240 | 30 | 270 |
| 18 | 231 | 38 | 269 |
| 19 | 129 | 100 | 229 |
| 20 | 0 | 115 | 115 |
| 21 | 170 | 33 | 203 |
| 22 | 29 | 29 | 58 |
| 23 | 76 | 9 | 85 |
| 24 | 61 | 39 | 100 |
| 25 | 43 | 25 | 68 |
| 27 | 76 | 15 | 91 |
| 28 | 77 | 35 | 112 |
| 29 | 0 | 127 | 127 |
| 30 | 187 | 36 | 223 |
| 31 | 176 | 28 | 204 |
| 32 | 131 | 23 | 154 |
| 33 | 154 | 43 | 197 |
| 35 | 118 | 51 | 169 |
| 36 | 40 | 14 | 54 |
| 37 | 38 | 8 | 46 |
| 38 | 161 | 29 | 190 |

| | | | |
|---|---|---|---|
| 39 | 105 | 15 | 120 |
| 40 | 135 | 42 | 177 |
| 41 | 152 | 73 | 225 |
| 42 | 185 | 12 | 197 |
| 43 | 35 | 209 | 244 |
| 44 | 0 | 93 | 93 |
| 45 | 48 | 44 | 92 |
| 46 | 230 | 38 | 268 |
| 47 | 173 | 29 | 202 |
| 48 | 189 | 25 | 214 |
| 49 | 60 | 18 | 78 |
| 50 | 147 | 17 | 164 |
| 51 | 213 | 17 | 230 |
| 52 | 176 | 32 | 208 |
| 53 | 118 | 49 | 167 |
| Total | 5882 | 1992 | 7874 |

The same procedure was applied for both attainment tests (Mathematics, English) and from these three together a new variable was derived that indicated "1"-no assessment; and "0"-at least one assessment in these three tests.

This analysis shows that respondents have either answered all items that were given to them (taken the different item numbers across version into account) or they have answered none of them. No specific coding of missing item data is necessary.

The variable *nosci* already provides information regarding whether someone who was supposed to do the science test did not do it, since all students should have done the science test. This is a bit more complicated for the Maths and English test, since not all students were supposed to do it. Therefore two new variables are used that can be used to check whether students supposed to do either test (coded in *GLMeasure*), did actually do it.

With the variable *nodata* panel attrition was assessed for students that had NPD data, since any student should actually should have provided at least one of these tests. Overall, $N = 1579$ students did not provide any post-intervention data. Of these, n = 495 (160 + 115 + 127 + 93) were due to schools 5, 20, 29 and 44 dropping out of the study.

A new variable was created within this sample to identify the reason for missingness and to present the descriptive statistics for these students

| | Students in Analysis | Students Missing | Students from Schools dropping out | Students without those where school dropped out |
|---|---|---|---|---|
| Female | 48.4% | 47.1% | 46.1 | 49.3% |
| Free School Meal | 27.9% | 35.7% | 37.8% | 31.1% |
| KS2 results | 28.71 (SD=4.19) | 27.89 (SD=4.61) | 27.79 (SD=4.68) | 28.12 (SD=4.46) |
| N | 6295 | 1579 | 1084 | 495 |

And finally, three variables were coded to represent the share of free school meals, share of females and the mean attainment level at the school in the Key Stage-2 assessments. These variables will be used as predictors in the selection analysis to test whether certain schools were more or less likely either to take part in this study themselves or their respective students.

*2) Randomisation of pre-intervention attainment levels*

The purpose of the selection analysis is twofold. First, four schools did not provide any data for the trial at all and it has to be assessed whether and how much these four schools differed on baseline variables from the other 49 to explore potential limitations of generalisability.

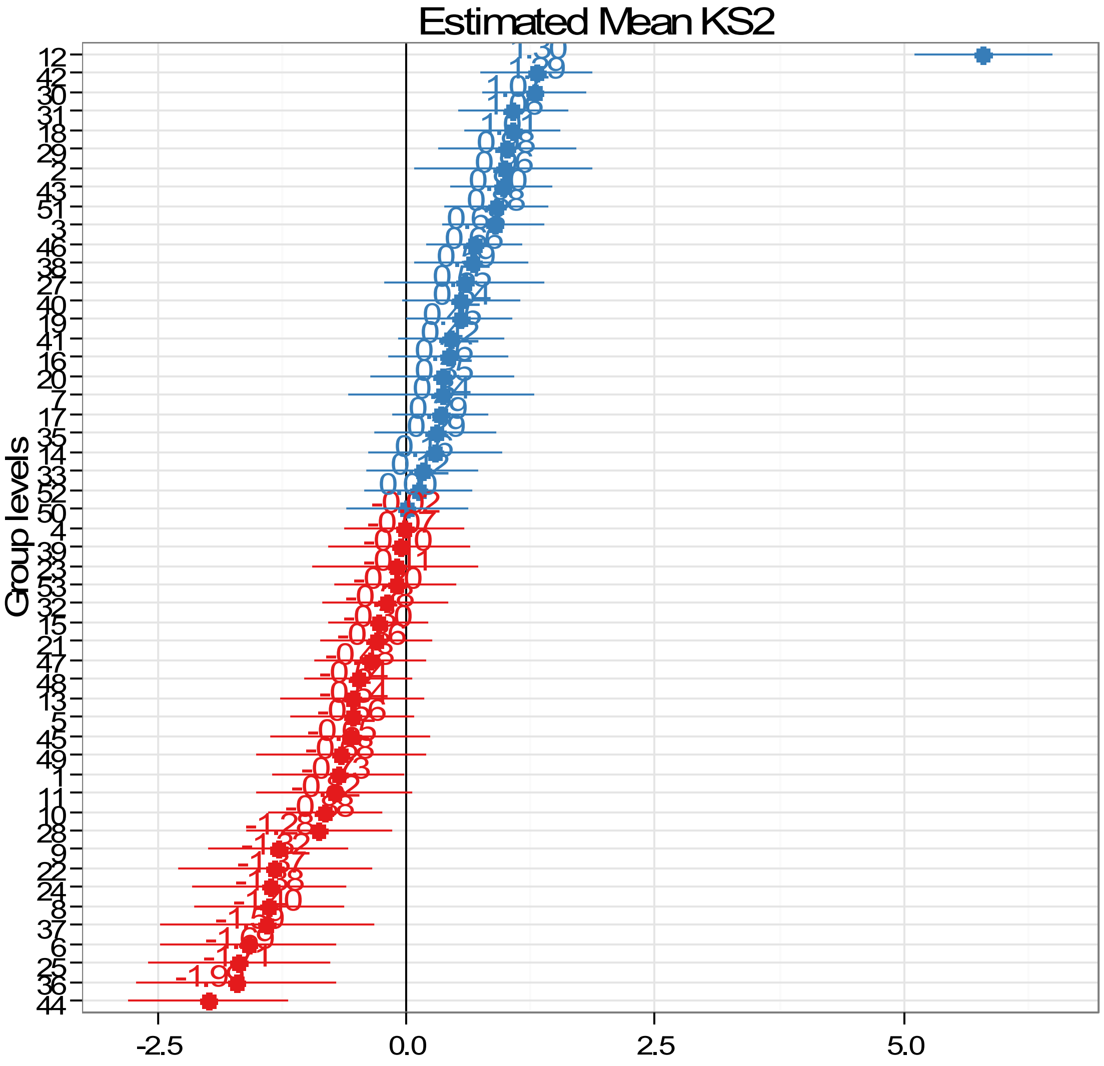
The second step will investigate within and between schools (multilevel model) whether there are systematic differences between schools regarding who actually took part in the second assessment after the intervention took place. The software used in this step is R 3.2.2[5].

We first look at the random effects and descriptive statistics for the schools that dropped out. Four schools were lost during the study: no. 5, 20, 29, and 44.

[5] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

The intraclass correlation for the Key Stage 2 results was relatively small with ICC = .08 [1.579/(1.579+17.189)=.084; $N$ = 7874; $n_{\text{school}}$ = 51]. The overall mean across schools was $M_{\text{KS2}}$ = 28.34 ($SE$ = .18). Figure 1 shows the caterpillar plot of the estimated random effects for the schools, which are the individual deviations from this overall mean (the "0" in figure 1). Of the schools that did not provide post-intervention data, school 44 is the one with the lowest estimated pre-study attainment out of the whole sample. The other three schools are within the overall distribution.

Figure 1: Distribution of KS2 intercepts for n=51 schools that provided NPD data



Adding *treatment* as a predictor to differentiate between the schools results in a non-significant effect (Type II Wald $\chi^2$-test; fixed effect = .046 ($SE$ = .36); $\chi^2_{\text{df=1}}$=1.56, $p$ = .21). This remained also true when controlling for other school level pre-intervention predictors (share of free school meals; share of gender).

With view to the available variables, pre-intervention attainment was not systematically correlated with the treatment variable and most of the variance in attainment was observed within schools and only a minor share of variance was due to between-school differences (*ICC*). Three of the four schools dropping out of the study were well within the distribution of the other schools, one was the school with the lowest estimated attainment in KS2.

<u>3) Selection analysis</u>

The data of the schools that entered the study are analysed for systematic selection effects regarding which students actually took part in the post-intervention tests. For this a generalised linear mixed model (GLMM) was run to predict whether a student within a given school provided any data or not (*nodata*, see above).

Running a GLMM with only a random effect for the schools reveals that quite a range of participation rates are observed. The estimated school odds ratios range from $OR_{40}$ = 0.33 (raw participation: 171 of 177) to $OR_{43}$ = 38.79 (raw participation: 36 of 249).

When school 43 is removed from the sample (different drop-out mechanism) and the model is re-run, it is possible to approximate the intraclass correlation (Skrondal & Rabe-Hesketh, 2005[6]): .5138 / (.5138 + pi²/3) = .136. The probability to drop out of the sample therefore does not strongly depend on differences between schools (and consequently all BIC comparisons indicate that adding a second level predictor does not enhance model fit). On the individual level only FSM ("yes" increases the probability overall to drop out) and KS2 (higher scores decrease the probability) predict the probability of dropping out of the sample. These two variables provide the best prediction of missingness in this sample with a random effect for FSM (i.e. different function per school) and a fixed effect for the KS2 score (same across schools, very little variance only).

The derived function is then applied to all schools in the participating sample. Treatment was kept although not significant, just to make sure that it was not a predictor (especially across the bootstrap samples drawn in the analyses). Variance of the weights was slightly smaller with treatment in it, but t-test showed that the weights were not significantly different between groups after 1/probab transformation.

The model is in then applied to all participating schools to produce individual student estimates for the probability of taking part in the study, which are then used as inverse probability weights (Seaman & White, 2013[7]). Since the main analysis for the primary outcome is a bootstrap-based analysis, the

[6] Rabe-Hesketh, S., & Skrondal, A. (2005). Multilevel longitudinal modeling using Stata. College Station, Texas: Stata Press.

[7] Seaman, S.R., & White, I.R. (2013). Review of inverse probability weighting for dealing with missing data. *Stat Meth Med Res*, *22*, 278-295.

weights will also be estimated each time since the sample composition is different for each bootstrap dataset.

**Data Analysis Tools**

- The data base for student records was built and maintained in Access.
- The data were transferred as an SPSS file from data management (LE) to data analysis (JRB); descriptive analysis for the definition of the different samples used in this analysis was performed on this original file (JRB).
- This SPSS file was read in via the `foreign`[8] package into R 3.2.2[9].
- Data management and descriptive analyses were performed with R's basic functional capabilities as well as the package `Rcmdr`[10] and the package `plyr`[11].
- Linear mixed models as well as generalised linear mixed models were performed with the package lme4[12]. Graphical model checking was done with the package `sjPlot`[13].
- The package `lavaan`[14] was used to determine the structural parameters for the secondary outcome analysis.
- Analysis of the clustered intervention effects was performed in Mplus 7.11 (Muthén & Muthén, 1998-2014[15]).
- For all exchanges between Mplus and R the package `MplusAutomation`[16] was employed.

[8] R Core Team (2015). foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, .... R package version 0.8-65. http://CRAN.R-project.org/package=foreign

[9] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[10] Fox, J. (2005). The R Commander: A Basic Statistics Graphical User Interface to R. Journal of Statistical Software, 14(9): 1--42.

[11] Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software, 40(1), 1-29. URL http://www.jstatsoft.org/v40/i01/.

[12] Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01.

[13] Lüdecke D (2015). _sjPlot: Data Visualization for Statistics in Social Science_. R package version 1.8.4, <URL: http://CRAN.R-project.org/package=sjPlot>.

[14] Yves Rosseel (2012). lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software, 48(2), 1-36. URL http://www.jstatsoft.org/v48/i02/.

[15] Muthén, L. K., & Muthén, B. O. (1998-2011). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

[16] Michael Hallquist and Joshua Wiley (2014). MplusAutomation: Automating Mplus Model Estimation and Interpretation. R package version 0.6-3. http://CRAN.R-project.org/package=MplusAutomation

**Data Analysis Primary Outcome**

*Descriptive Statistics*

Table 4 presents the descriptive statistics for the four tests used in this study. The mean and median of all four scores are very close to each other, indicating fairly symmetric distributions. Observed minima and maxima, as well as the standard deviations, indicate that a broad range of scores was realised in each of the measures in this study.

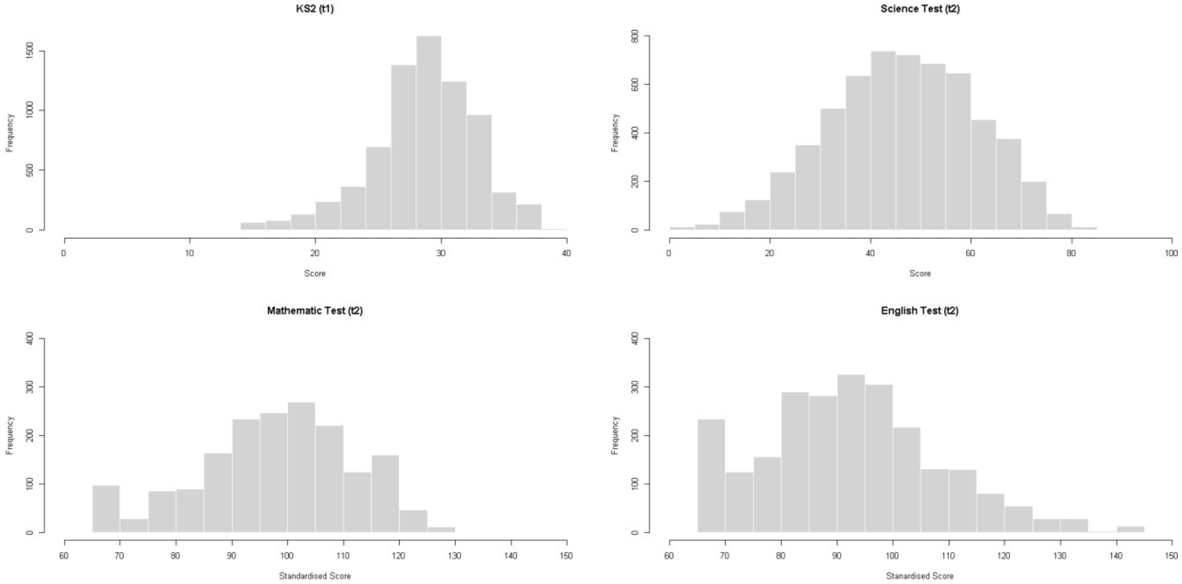
Table 4. Descriptive statistics for the tests used in this study

| | Science Test (t2) | KS2 (t1) | Mathematic (t2) | English (t2) |
|---|---|---|---|---|
| Mean, Median | 46.82, 47 | 28.57, 28.73 | 98.41, 99 | 93.01, 92 |
| SD | 14.75 | 4.28 | 13.67 | 15.49 |
| Min | 2 | 3 | 69 | 69 |
| Max | 85 | 39 | 129 | 129 |
| $N_{\text{Students}}$ | 5882 | 7379 | 1775 | 2397 |

Figure 1 presents histograms for all four tests, corresponding to the numbers presented in table 4. Although only one of the four distributions is close to a normal distribution (Science Test), all distributions are unimodal and corroborate the previously noted conclusion of fairly symmetric distributions (similar means and medians). Only for the English and Mathematics Test slight floor effects were observed with increased frequencies of the lowest possible score.

Overall the tests seem to be appropriately targeted for the population of schools and students. To address the non-normality in distributions a bootstrapping approach for model estimation was chosen (see below).

Figure 1. Histograms for the four tests



*Assessing Potential Marker Effects*

The primary outcome defined for this study was the result in the science test. To assess the reliability of the marking of this test, correlation between two independent markers as well as the percentage of total variance due to markers were determined. During the project, $N = 593$ randomly selected science tests were marked by two markers. The raw correlation between these two assessments was $r = .993$ (one-sided t-test: $t = 201.37$; $df = 591$; $p < .001$).

For a more detailed assessment of how much variability of the assessments was due to the marker and how much was due to differences between students, a generalisability analysis was done. A random effects model with three independent components (student, school, marker) was run (results in table 5)[17]. The results reveal that 78.3 % of the observed variance in scores is due to differences in scores, 20.8 % is due to differences between schools and less than a percent is due to the markers and the residual term, with the variation in markers being responsible for 0.3 % of variance. It is therefore safe to say that the assessment of the science test was independent of the marker who performed this assessment.

[17] Brennan, R.L. (2001). Generalizability theory. New York: Springer.

Table 5. Table of variance components to assess marker effects (*N* = 593)

| | Variance Component | Percentage of Total Variance |
|---|---|---|
| Student | 190.67 | 78.3 |
| School | 50.72 | 20.8 |
| Marker | 0.69 | 0.3 |
| Residual | 1.44 | 0.6 |

*Assessing the effect of the intervention*

To assess the size of potential effects of clustering on the results, standard intra-class correlations were estimated for the Science test as the main outcome as well as the aggregated Key Stage 2 (KS2) results (main control variable) and for the secondary outcomes (Mathematics and English tests). The ICCs are reported in table 6, all being of moderate size as expected in educational settings, the lowest achieved in the Mathematics assessment and the highest in the English test.

Table 6. Variance components and intraclass correlations for the three outcome measures and the control variable (KS2 from baseline)

| | Science Test (t2) | KS2 (t1) | Mathematic (t2) | English (t2) |
|---|---|---|---|---|
| Variance, Schools | 29.12 | 1.577 | 9.98 | 39.14 |
| Variance, Residual | 196.47 | 17.091 | 179.08 | 211.43 |
| ICC | .129 | 0.084 | .053 | .156 |
| $N_{schools}$ | 47 | 47 | 17 | 20 |
| $N_{students}$ | 5882 | 7379 | 1775 | 2397 |

Before the analysis, it was checked whether the available baseline characteristics were balanced across treatment group. For this, we estimated (generalised) linear mixed models predicting missingness of data and gender (generalised linear mixed model) and KS2 results (linear mixed model) which estimated random effects on school level and predicted these random effects with the treatment group. Treatment was not a significant predictor of missingness of data (logit estimate = .16, $p$ = .561; $N_{students}$ = 7379; $N_{schools}$ = 47). The same was true of gender (logit estimate = -.09, $p$ = .467; $N$ = 6926, ; $N_{schools}$ = 47, ITT population). The generalised linear mixed model for KS2 results did also not reveal any pre-study differences (estimate = .27, $p$ = .475; $N$ = 6926; $N_{schools}$ = 47, ITT population).

To assess the effect of the intervention, a multilevel model was run in Mplus. Within schools each individual's science score $SCI_{ij}$ was corrected for gender and previous attainment (averaged KS2 results as provided in the NPD data set). The KS2 scores were grand mean centered ($KS2c_{ij}$), this way each school's intercept ($\beta_{0j}$) represents the expected science achievement for one of their respective students with performance equivalent of the overall KS2 mean. If the intervention had an effect in those schools where it was delivered compared to those where it has not been delivered, the mean science achievement for these school-level average students should be expected to be different (by $\gamma_{00}$, see formula 1). The statistical analysis tests whether this parameter is larger than "0", employing a conventional significance level of $p \leq .05$ (bootstrapped, $b$ = 1000). If this fixed effect is positive and significant, this means that in schools receiving the intervention science attainment was increased compared to those not receiving the intervention, and the LTSS program had a positive effect on science attainment.

$$\begin{aligned} SCI_{ij} &= \beta_{0j} + \beta_{1j} gender_{ij} + \beta_{2j} KS2c_{ij} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} Treatment_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{10} \\ \beta_{2j} &= \gamma_{20} + u_{20} \end{aligned}$$

Formula 1

The bootstrap is performed on school level, ie, within every school a sample is drawn with replacement and of the same size as the original sample for this school. These samples are then aggregated across schools and the mixed effects model as described above was estimated and the parameter $\gamma_{01}$, which is the expected mean difference of performance in the science test after controlling for KS2 and gender, was saved for every run.

As a first sensitivity analysis, this multilevel model was re-estimated on the same sample, but this time the students were weighted for their probability to have no observation on the follow-up assessment. Since only very few characteristics were available to determine the probability of a respondent not providing data at the post-intervention assessment, a simple weighting model was estimated instead of using imputation or other approaches. In every run a generalised mixed model predicting the probability of dropping out before the final assessment was calculated. On the original data, models with increasing number of predictors and complexity were estimated and two predictors were identified via information criteria (BIC) as being most likely to provide relevant information on missingness: Whether the respective student was ever eligible for free school meals (if yes, they were more likely to have dropped out of the sample; large variance between schools therefore a random effect) and the KS2 result (the higher the score, the less likely the student was to have dropped out; nearly no variance between schools, therefore entered only as a fixed effect). *Treatment* was added as a second level predictor for school means since the weights should not influenced by treatment related variation in drop-out rates, although overall it was not a relevant or significant predictor. This model is fit to every bootstrap sample and the predicted probability for each individual student is used to weigh their observations in a repeat run of the above mixed model predicting the Science Test results. School 43 is not used to estimate the weights, since this school misunderstood for which classes they should have send the data, but the prediction of weights is applied to all participating schools (fixed effects only out-of-sample prediction for school 43). Any analysis that is described as *weighted* in this report used these weights for the probability of not providing data at the follow-up assessment.

Table 7: Fixed effect regression coefficients and their respective random effects from the multilevel analysis for the primary outcome (Science Test), with empirical single run coefficients and their respective normal-theory confidence intervals in square brackets.

| | Unweighted Analysis | Weighted Analysis |
|---|---|---|
| Fixed Effects | | |
| Intercept, $\gamma_{00}$ | 46.42<br>[44.92, 47.93] | 48.93<br>[47.37, 50.49] |
| KS2 (centred on grand mean level), $\gamma_{20}$ | 2.57<br>[2.48, 2.66] | 2.64<br>[2.55, 2.73] |
| Gender, $\gamma_{10}$ | .40<br>[-.27, 1.07] | .59<br>[-.12, 1.30] |
| Treatment, $\gamma_{01}$ | -.13<br>[-2.47, 2.22] | -.18<br>[-2.55, 2.19] |
| | | |
| Random Effects | | |
| Intercept, $u_{00}$ | 15.21<br>[7.85, 22.57] | 15.66<br>[8.26, 23.05] |
| Gender, $u_{10}$ | 1.84<br>[.33, 3.34] | 2.50<br>[.81, 4.31] |
| KS2 (centered on grand mean), $u_{20}$ | .04<br>[.01, .08] | .04<br>[.01, .07] |
| | | |
| Residual | 88.73<br>[82.91, 94.54] | 86.56<br>[80.87, 92.24] |

Note. Model estimated in Mplus 7.11, restricted maximum likelihood estimator; all regression coefficients are unstandardised estimates; $N_{\text{school}} = 47$, $N_{\text{student}} = 5882$; Intraclass correlation of the Science Test across bootstrap samples: $M_{\text{ICC}} = .14$ [.11; .16]

Of the control variables, gender was not a significant predictor of Science Test attainment, neither in the weighted nor the unweighted analyses ($\gamma_{10}$ non-significant). Nevertheless, as the estimates of the random effect for gender across schools show ($u_{10}$), there was considerable between-school variance. The available proxy for pre-intervention performance (KS2 results), was highly significant, with students performing better in KS2 also performing better in the Science Test (see also figure below for individual school regression lines).

The effect for the treatment variable was estimated as $\gamma_{01} = -.13$ (weighted: $\gamma_{01} = -.18$), which means that the estimated difference between mean Science Test attainment for schools receiving the intervention and those, not receiving the intervention was equal to -.13 points. This difference was not significant, neither in the weighted nor in the unweighted analysis. The EEF's policy on the analysis of evaluation studies suggests calculating as an effect size the treatment effect divided by the total variance. The estimated coefficient $\gamma_{01}$ for the treatment variable is the estimate of the difference between intervention and control group, adjusted for KS2 performance and gender. Therefore, the effect sizes (ES) from weighted and unweighted analyses are[18]:

[18] As derived from the suggested formula on page 5 of the "Policy on Analysis for EEF evaluations", Educational Endowment Foundation, 30.10.2015. Since only one intervention was delivered in a school no interaction between school and intervention can be estimated.

$$ES = \frac{\left(\overline{Y}_T - \overline{Y}_C\right)_{adjusted}}{\sqrt{\sigma^2{}_s + \sigma^2{}_{error}}} = \frac{\gamma_{01}}{\sqrt{e_{ij} + u_{00}}}$$

$$ES_{\text{unweighted}} = \frac{-.13}{\sqrt{88.73 + 15.21}} = -.013 \qquad \text{Formula 2}$$

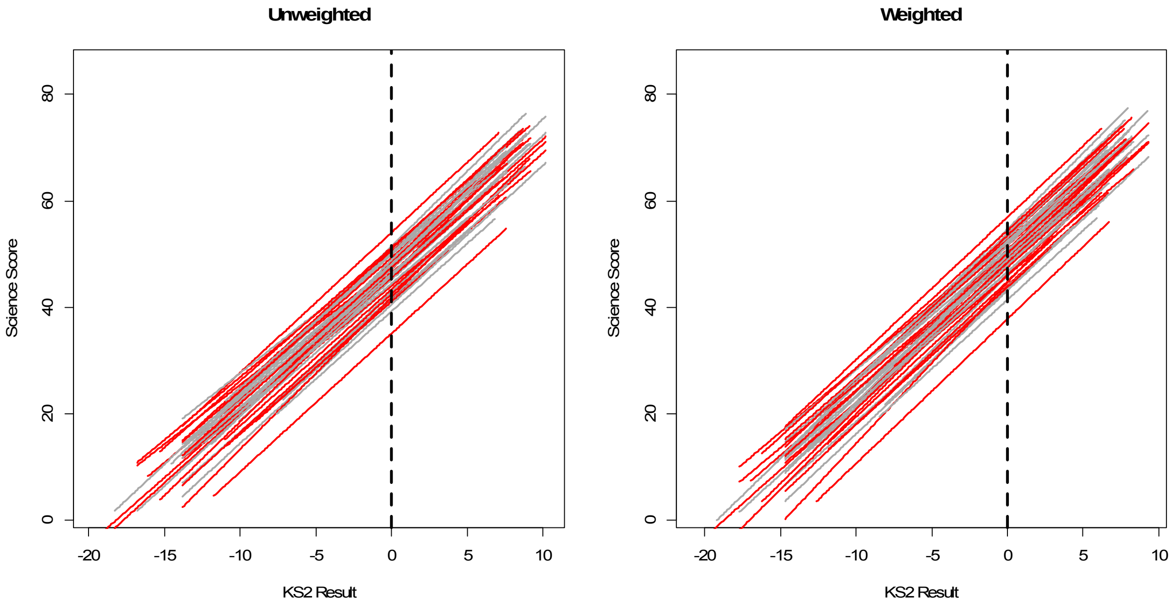
$$ES_{\text{weighted}} = \frac{-.18}{\sqrt{86.56 + 15.66}} = -.018$$

This effect size can be classified as very small and the bootstrapped confidence interval for this effect size ranges from -.091 to .044 (weighted: -.104 to .044).

Based on these estimates it is possible to perform a post-study power estimation, whether it would have been possible to find a significant small effect (*ES* = .20) with exactly this design and data gathered. For this purpose, the fixed effect was calculated that would have been necessary to result in this effect size [$\gamma_{01.\text{power}}$ = 0.2*(sqrt(88.73+15.21)) = 2.04], i.e. schools in the intervention would have needed to have about 2.04 science test points higher mean scores than the control schools. The post-study evaluation of power through a simulation study allows using exactly the estimates as they were obtained in the current study and only change those parameters that are relevant for this step. It also uses the exact same numbers of students per school instead of a mean value, which captures the differential precision across schools better than generic algorithms. In our evaluation we used therefore all parameters as presented in table 7 and only changed the estimate of $\gamma_{01}$ to the calculated value that would be equivalent to an effect size of d = 0.20. We used the same number of schools (47) with the same specific number of respondents as realised in the study within each school, total $N_{\text{power}}$ = 5882 and the same overall male/female ratio. Schools were simulated as being randomly allocated to the LTSS intervention.

Across *b* = 1000 runs (run in Mplus 7.11), an effect that would have been equal to a controlled between-school effect size of *d* = .20 was found to be significant at *p* < 5% in 42.2% of the simulated samples (i.e., power = .422).

To ease evaluation of the empirical results, the following figure displays the regression lines that are estimated for every school in the sample by this model. Each regression line runs from the lowest observed value within that school to the highest observed value. The regression lines represent the relationship between the KS2 results and the Science Test in each school. The grey lines represent schools from the control sample, red lines those from the intervention sample. The dashed line at *KS2* = 0 represents the overall mean in the full sample at which the schools from both samples should perform differently, if the hypothesis was true. As can be seen, the regression lines are overlapping and cover a similar range at this position on the x-axis. The variation in intercepts, i.e. points on the dashed zero-line, is a bit larger for the schools in the intervention sample, with both the highest performing as well as the lowest performing school in that sample.



The following figure illustrates just the variance for the intercepts across the two groups in this sample. Both panels present the density plots for the estimated means for all 47 schools in the main analysis (weighted and unweighted analysis, respectively; intervention = red, control = grey). Clearly, no shift of the central tendency is observed, neither in the weighted, nor in the unweighted analysis. If the treatment would have been effective, a shift of the red density function to the right should have occurred.

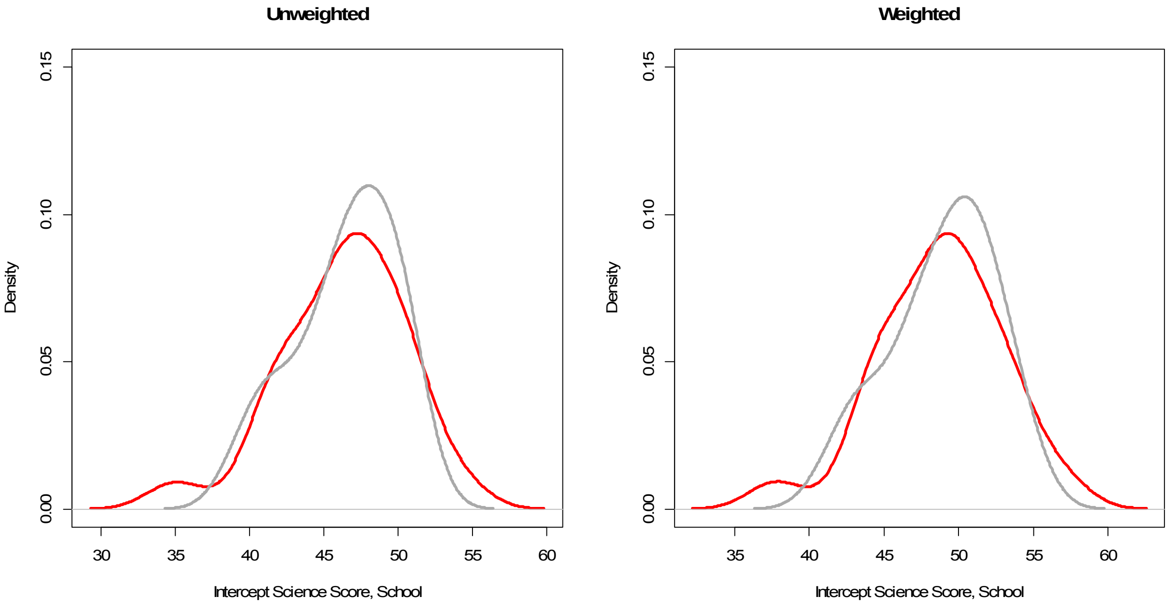


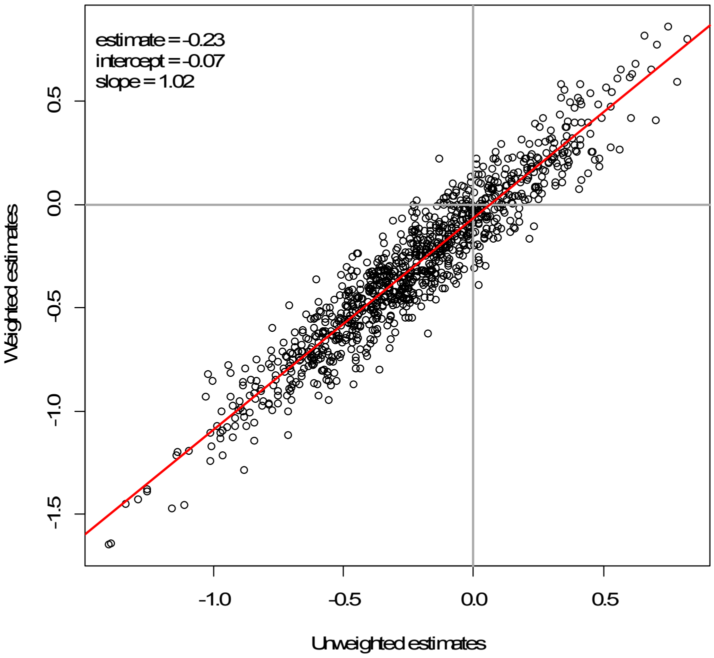
Table 7 above presented normal-theory based estimates and confidence intervals. Using the results from the bootstrap analyses did not change the conclusion regarding the effect of the treatment as can be seen from the results presented in table 8. Both the weighted and unweighted analyses (intention to treat) on the main sample resulted in small and non-significant estimates and the bootstrap confidence intervals included "0". The probability of obtaining a coefficient that is larger than "0" (i.e., in line with the hypothesis rejecting the NULL hypothesis) is $p$ = .254, far lower than to be expected for conventional significance levels.

Table 8 also reports the estimates for several sensitivity analyses of this effect, taking into account breaches of the randomisation protocol. Since not all schools did deliver the treatment they were randomly assigned to, in the *per protocol* analysis only those schools that delivered the intervention they were randomised to were analysed (only if LTSS was delivered when randomised to LTSS; only if control was delivered when randomised to control), excluding schools that violated the randomisation. In this sample, as in the ITT sample, small and non-significant effect estimates were found (fourth and fifth column of table 8).

Additionally, we also tested the effect of the intervention actually received, i.e. we analysed the data based on the intervention that was actually delivered ("as is"), regardless of which treatment they were randomised to. In practice, this involved re-allocating students in three schools where we knew that, although they were in the intervention group, they had not received the programme. Again small and non-significant effect estimates resulted from this analysis (sixth and seventh column of table 4). All confidence intervals include "0" and the probability of a replicated coefficient larger than "0" (i.e., in line with the hypothesis rejecting the NULL hypothesis) never reaches conventional levels of significance. Although these two latter tests should give the intervention a greater chance to show an effect (albeit usually due to then systematic differences other than the pure intervention effect), it fails to do so.

Table 8. Bootstrap results for the fixed effect estimate of the treatment from the main analysis as well as all sensitivity analyses

| | ITT, unweighted Analysis | ITT, weighted Analysis | Per Protocol Analysis | Per Protocol Weighted | "As is"[19] | "As is", weighted[20] |
|---|---|---|---|---|---|---|
| Treatment Mean | -.227 | -.299 | .067 | -.012 | .292 | .230 |
| Treatment Median | -.227 | -.313 | .049 | -.028 | .228 | .158 |
| Lower 5%-bound | -.835 | -.923 | -.549 | -.669 | -.495 | -.738 |
| Percentage of coefficients > 0 | .254 | .216 | .563 | .476 | .659 | .627 |



There is some variance in the estimates as table 8 illustrates. For the following figure the weighted estimates for the treatment effect were regressed on the unweighted ones to assess potential non-linear differences. As is clearly seen, the weighting does not lead to differences between the solutions apart from the shift of the treatment being estimated overall as less efficacious under the weighting scheme controlling for the probability of missing follow-up data.

*SUMMARY*

These results indicate that (under various different assumptions and within different subsamples that were used as sensitivity analyses) the LTSS intervention did not have an effect on science attainment.

[19] Since three schools delivered both treatment and control, this analysis is only a mixed linear regression controlling for school effects, but treating "Treatment" as a within school variable.

[20] Since three schools delivered both treatment and control, this analysis is only a mixed linear regression controlling for school effects, but treating "Treatment" as a within school variable.

*Interaction effects with other variables*

In search of potential subgroup effects, the relevance of interaction effects with the treatment variable were assessed. Pre-specified interaction terms were free school meals (FSM), gender, and pre-intervention attainment. For each of these variables a cross-level interaction effect will be defined in the following manner (example gender):

$$\begin{aligned} SCI_{ij} &= \beta_{0j} + \beta_{1j} gender_{ij} + \beta_{2j} KS2c_{ij} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} Treatment_j + u_{00} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} Treatment_j + u_{10} \\ \beta_{2j} &= \gamma_{20} + u_{20} \end{aligned} \qquad \text{Formula 3}$$

To test whether there was a relevant interaction term between gender and the LTSS intervention (and subgroup analyses would therefore be relevant), *Treatment* is added as a second-level predictor of the relationship between gender and the Science Test result in each school. The coefficient $\gamma_{11}$ was bootstrapped using the same samples as described above and central 95%-confidence interval. The same procedure is applied for FSM and KS2 results, but for the latter this will be tested both for continuous KS2 scores as well as KS2 terciles (calculated on the full sample). These analyses will be conducted weighted as well as unweighted for missingness.

Table 9 presents the results based on the 95%-confidence intervals (bootstrapped) for weighted and unweighted analyses. There was no interaction effect for KS2 results as a continuous variable (second and third column). When the KS2 results are grouped into terciles (full sample) then an interaction effect is observed for the third (top) tercile under weighted and unweighted analyses (fourth and sixth columns; see below). Students in the top tercile of the KS2 pre-test in the treatment group do about 1.42 (unweighted; 1.56 weighted) points worse in the science test compared to a student with all other variables being equal from the first tercile (reference group). This effect is followed up below with a subgroup analysis.

An interaction effect for gender is found as well. Females in the treatment group seem to do better (by about 1.1 points in the Science Test) than male students in the control group. This effect is followed up below with a subgroup analysis.

The confidence interval for the FSM X Treatment interaction includes 0 in both analyses, meaning that neither of these interaction effects is statistically significant.

Table 9: Results for pre-defined interaction analyses

| | KS2 | KS2 weighted | KS2: second tercile vs. first | KS2 third tercile vs. first | KS2 second tercile vs. first; weighted | KS2 third tercile vs. first; weighted | Gender | Gender, weighted | FSM | FSM weighted |
|---|---|---|---|---|---|---|---|---|---|---|
| Interaction Term, mean | -.082 | -.062 | -.333 | -1.425 | -.486 | -1.561 | 1.167 | 1.083 | -.215 | -.621 |
| Interaction Term, median | -.082 | -.062 | -.335 | -1.422 | -.495 | -1.564 | 1.151 | 1.070 | -.216 | -.648 |
| Bootstrap Confidence Interval | [-.205, .057] | [-.195, .080] | [-1.717, .999] | [-2.774, -.068] | [-1.903, .898] | [-3.071, -.126] | [.258, 2.157] | [.133, 2.127] | [-1.429, .975] | [-1.907, .632] |

To follow the effect of the top KS2 tercile ("bin") up, table 10 presents the results for the analyses within the first two terciles only (left column) and within the third tercile only (right column). The intervention effect is not significant in either of them and again indicating only very small effect sizes.

Table 10: Follow up Analyses for lower two terciles vs. top tercile of KS2 results (pre-test)

| | First and second tercile<br>Unweighted Analysis[21] | Third tercile<br>Unweighted Analysis[22] |
|---|---|---|
| Fixed Effects | | |
| Intercept, $\gamma_{00}$ | 33.32<br>[31.96, 34.67] | 58.50<br>[56.66, 60.34] |
| KS2 tercile 2 (tercile 1 as reference), $\gamma_{20}$ | 12.01<br>[10.57, 13.45] | -- |
| Gender, $\gamma_{10}$ | 1.11<br>[-.30, 1.14] | .22<br>[-.72, 1.17] |
| Treatment, $\gamma_{01}$ | .64<br>[-1.59, 2.87] | -.39<br>[-2.94, 2.16] |
| KS2 tercile 2 (tercile 1 as reference) | -.33<br>[-2.23, 1.58] | -- |
| | | |
| Random Effects | | |
| Intercept, $u_{00}$ | 11.50<br>[4.21, 18.78] | 15.58<br>[8.52, 22.64] |
| Gender, $u_{10}$ | .51<br>[-1.56, 2.57] | 1.00<br>[-1.44, 3.45] |
| KS2 tercile 2 (tercile 1 as reference), $u_{20}$ | 5.34<br>[1.97, 8.70] | -- |
| | | |
| Residual | 110.40 | 95.54 |

[21] Since it is unclear how this subgroup selection relates to the dropout mechanism, the analysis was only conducted in an unweighted fashion.

[22] Since it is unclear how this subgroup selection relates to the dropout mechanism, the analysis was only conducted in an unweighted fashion.

| | | |
|---|---|---|
| | [103.60, 117.20] | [86.55, 104.53] |
| | $N_{\text{school}} = 46$, $N_{\text{student}} = 3837$ | $N_{\text{school}} = 47$, $N_{\text{student}} = 2045$ |

Note. Model estimated in Mplus 7.11, restricted maximum likelihood estimator; all regression coefficients are unstandardised estimates

The corresponding effect sizes are:

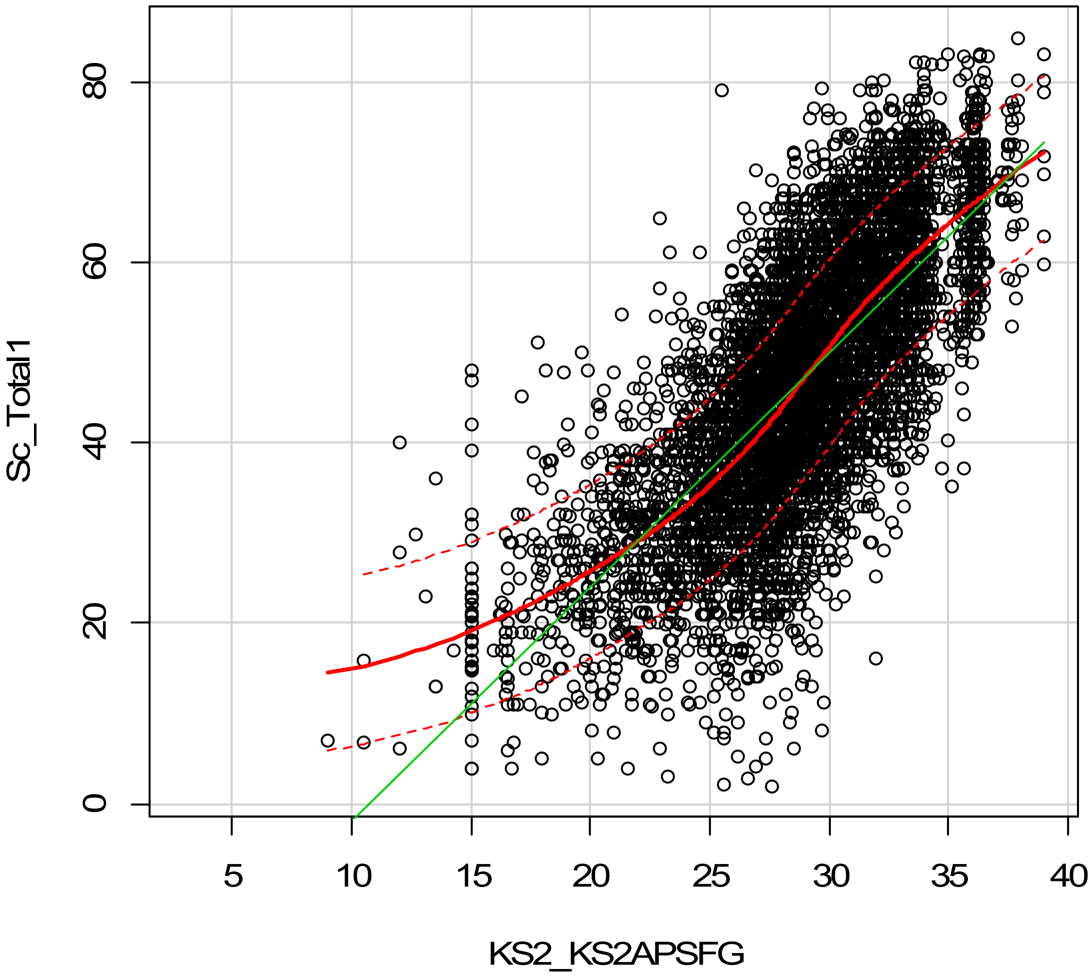
$$ES = \frac{\left(\overline{Y}_T - \overline{Y}_C\right)_{adjusted}}{\sqrt{\sigma^2{}_s + \sigma^2{}_{error}}} = \frac{\gamma_{01}}{\sqrt{e_{ij} + u_{00}}}$$

$$ES_{\text{tercile1+2}} = \frac{.64}{\sqrt{110.40 + 11.50}} = .059$$

$$ES_{\text{tercile3}} = \frac{-.39}{\sqrt{95.54 + 15.58}} = -.037$$

Looking at the main effect of the pre-test, a student who starts with a very high pre-test will do on average much better in the Science Test than a student in the lower attainment groups. The marginal expected value for a student in the lowest tercile is 33.32 points in the science test (intercept of second column in table 10); the marginal expected increase for a student from the second tercile is an additional 12.01 points ($\gamma_{20}$ in table 10; ~45.33 points). The expectation for a student in the top tercile is then 58.50 points. While this seems to be a relatively linear increase, the treatment effect of the LTSS intervention is positive for students in the lower two percentiles (an .64 additional points in the Science Test for students in the LTSS intervention), this effect is negative for the students in the top percentile (a decrease of .39 points in the LTSS intervention group). While both intervention effects remain non-significant (confidence intervals include 0 and LTSS has neither a positive nor a negative effect in either group) and small in size (see calculations above), their difference might be enough to lead to the statistical significance of the interaction effect.

An additional explanation for this effect might be that the KS2 results are not linearly related in the top segment of pre-test scores. As the following figure depicts for the ITT sample, the relationship between the two tests is attenuated at the top end, which leads to a smaller increase in Science Scores than expected by the linear trend in the middle of the distribution. This effect might be correlated with the allocation to the intervention conditions (remember: top scoring school is in the intervention group and all students in that school score in the top tercile of the KS2 distribution).



To follow the effect of gender up, table 11 presents the results for the analyses within males (left column) and within females only (right column). The intervention effect is not significant in either of them and again indicating only very small effect sizes.

Table 11: Follow-up analyses for gender

| | Males<br>Unweighted Analysis[23] | Females<br>Unweighted Analysis[24] |
|---|---|---|
| Fixed Effects | | |
| Intercept, $\gamma_{00}$ | 46.43<br>[44.85, 48.01] | 46.23<br>[44.87, 47.60] |
| KS2 (centered on grand mean), $\gamma_{20}$ | 2.51<br>[2.40, 2.62] | 2.66<br>[2.56, 2.77] |
| Treatment, $\gamma_{01}$ | -.55<br>[-3.01, 1.91] | .94<br>[-1.32, 3.19] |
| | | |
| Random Effects | | |
| Intercept, $u_{00}$ | 16.22<br>[8.48, 23.97] | 13.22<br>[4.90, 21.54] |
| KS2 (centered on grand mean), $u_{20}$ | .04<br>[-.02, .11] | .04<br>[-.001, .07] |
| | | |
| Residual | 102.40<br>[93.85, 110.94] | 73.91<br>[68.95, 78.88] |
| | $N_{\text{school}} = 47$, $N_{\text{student}} = 3041$ | $N_{\text{school}} = 46$, $N_{\text{student}} = 2841$ |

Note. Model estimated in Mplus 7.11, restricted maximum likelihood estimator; all regression coefficients are unstandardised estimates

The corresponding effect sizes are:

$$ES = \frac{\left(\overline{Y}_T - \overline{Y}_C\right)_{adjusted}}{\sqrt{\sigma^2{}_s + \sigma^2{}_{error}}} = \frac{\gamma_{01}}{\sqrt{e_{ij} + u_{00}}}$$

$$ES_{\text{males}} = \frac{-.55}{\sqrt{102.40 + 16.22}} = -.050$$

$$ES_{\text{females}} = \frac{.94}{\sqrt{73.91 + 13.22}} = .100$$

Overall, male and female students reach the same marginal outcome (a score of ~46 points; see intercepts for both groups in table 7). A male student in an LTSS school will do slightly worse (marginal loss: -.55 points in the Science Test) than a male student in a control school. A female student is estimated to do slightly better in an LTSS school (a gain of ~.94 points in the Science Test) than a female student in a control school. While both intervention effects remain non-significant (confidence intervals include 0 and LTSS has neither a positive nor a negative effect in either group) and small in size (see calculations above), their difference might be enough to lead to the statistical significance of the interaction effect.

[23] Since it is unclear how this subgroup selection relates to the dropout mechanism, the analysis was only conducted in an unweighted fashion.

[24] Since it is unclear how this subgroup selection relates to the dropout mechanism, the analysis was only conducted in an unweighted fashion.

*Conclusion*

While interaction effects were detected between treatment and (a) scoring in the top tercile of the pre-test distribution and (b) gender, subgroup analyses revealed that the treatment did not have a statistically significant effect in any of the subgroups and effect sizes were small (and smaller than the study was planned for). This means that LTSS does not seem to have a differential effect depending on gender or prior attainment.

*Report for Results on FSM-only*

The analysis for the FSM-only population also shows that the treatment was unlikely to have an effect here. The effect size is again very small.

$$ES = \frac{\left(\overline{Y}_T - \overline{Y}_C\right)_{adjusted}}{\sqrt{\sigma^2{}_s + \sigma^2{}_{error}}} = \frac{\gamma_{01}}{\sqrt{e_{ij} + u_{00}}}$$

$$ES_{\text{unweighted}} = \frac{-.29}{\sqrt{92.05 + 13.76}} = -.028$$

This effect size can be classified as very small and the bootstrapped confidence interval for this effect size ranges from -.178 to .065. Table 12 presents the estimates from the analysis including the confidence interval that includes 0, i.e. showing the non-significant effect.

Table 12: Results of the FSM only analysis

| | Unweighted Analysis[25] |
|---|---|
| Fixed Effects | |
| Intercept, $\gamma_{00}$ | 41.44<br>[39.80, 43.03] |
| KS2 (centred on school level), $\gamma_{20}$ | 2.47<br>[2.33, 2.61] |
| Gender, $\gamma_{10}$ | .17<br>[-1.05, 1.39] |
| Treatment, $\gamma_{01}$ | -.29<br>[-2.79, 2.21] |
| | |
| Random Effects | |
| Intercept, $u_{00}$ | 13.76<br>[3.87, 23.65] |
| Gender, $u_{10}$ | 2.88<br>[-2.52, 8.27] |
| KS2 (centered on grand mean), $u_{20}$ | .05<br>[-.03, .13] |
| | |
| Residual | 92.05<br>[83.59, 100.51] |

Note. Model estimated in Mplus 7.11, restricted maximum likelihood estimator; all regression coefficients are unstandardised estimates; $N_{\text{school}} = 47$, $N_{\text{student}} = 1608$

[25] Since it is unclear how this subgroup selection relates to the dropout mechanism, the analysis was only conducted in an unweighted fashion.

## SECONDARY OUTCOME

In addition to the analysis of the primary outcome, the PIM (GL Progress in Maths) and PTE (GL Progress Test in English) were collected. To investigate whether differential and/or positive transfer effects can be detected across the different domains that these tests assess the original plan was to extend the multilevel model from the Primary Outcome Analysis. Since the overlapping assessment plan originally planned for in this study could not be realised, it was decided later on (based on feedback from the EEF), that only the schools would be analysed that received either the PIM or the PTE in separate sets of analyses. Although this came with the disadvantages (a) of reducing the sample size for the follow-up analysis and thereby reducing the statistical power as well as (b) the need to control the statistical significance level for multiple testing (set to $p$ = .025), this seemed to be the most appropriate strategy.

The same statistical model was run as for the primary outcome, but in addition it was controlled for on school level whether the PIM and PTE were administered online or as paper-pencil versions. Again, a positive effect of the LTSS intervention on both of these outcomes will be seen as further corroboration that the LTSS intervention had a positive effect on Science attainment and in addition positive transfer effects on Mathematical and English Attainment could be observed.

Table 13 presents the estimates from single normal-theory runs for each outcome. *Gender* and *KS2* results are again used as within-school control variables and *Treatment* is the predictor for the effect of the treatment on both. Maths and English tests are further corrected on between-school level for online vs. paper assessment.

As the confidence intervals for both unweighted and weighted analyses of the English Test in table 13 show, the estimated coefficients are negative, but non-significant based on normal-theory confidence intervals. Predictors of English attainment in this sample are the KS2 results (per point increase in KS2 a gain of 2.5 points in English) and *gender* (female students doing slightly more than 3 points better than male students).

Table 13: Fixed effect regression coefficients and their respective random effects from the multilevel analysis for the secondary outcome English Test, with empirical single run coefficients and their respective normal-theory confidence intervals in square brackets.

| | PTE<br>Unweighted Analysis | PTE<br>Weighted Analysis |
|---|---|---|
| Fixed Effects | | |
| Intercept, $\gamma_{00}$ | 94.01<br>[90.35, 97.66] | 96.71<br>[92.99, 100.43] |
| KS2 (centered on grand mean), $\gamma_{20}$ | 2.53<br>[2.32, 2.73] | 2.76<br>[2.54, 2.98] |
| Gender, $\gamma_{10}$ | 3.22<br>[1.97, 4.48] | 3.66<br>[2.22, 5.09] |
| Treatment, $\gamma_{01}$ | -1.68<br>[-5.44, 2.07] | -1.56<br>[-5.30, 2.19] |
| Online, $\gamma_{02}$ | -3.64<br>[-7.88, .61] | -4.09<br>[-8.36, .188] |
| | | |
| Random Effects | | |
| Intercept, $u_{00}$ | 16.03<br>[7.63, 24.43] | 16.55<br>[7.83, 25.26] |
| | | |
| Residual | 105.44<br>[93.44, 117.45] | 109.60<br>[96.97, 122.23] |
| | $N_{\text{school}} = 20$, $N_{\text{student}} = 2397$ | $N_{\text{school}} = 20$, $N_{\text{student}} = 2397$ |

Note. Model estimated in Mplus 7.11, restricted maximum likelihood estimator; all regression coefficients are unstandardised estimates

As the confidence intervals for both unweighted and weighted analyses of the Maths Test in table 14 show, the estimated coefficients are negative, but non-significant based on normal-theory confidence intervals. Predictors of Maths attainment in this sample are the KS2 results (per point increase in KS2 a gain of 2.5 points in Maths) and controlling for online seemed relevant in this case (those using PC administration scoring about 3.83 (unweighted) points higher in the Maths test).

Table 14: Fixed effect regression coefficients and their respective random effects from the multilevel analysis for the secondary outcome Maths Test, with empirical single run coefficients and their respective normal-theory confidence intervals in square brackets.

| | PIM<br>Unweighted Analysis | PIM<br>Weighted Analysis |
|---|---|---|
| Fixed Effects | | |
| Intercept, $\gamma_{00}$ | 96.41<br>[94.62, 98.20] | 99.07<br>[97.16, 100.97] |
| KS2 (centered on grand mean), $\gamma_{20}$ | 2.50<br>[2.36, 2.64] | 2.61<br>[2.47, 2.76] |
| Gender, $\gamma_{10}$ | -.40<br>[-1.22, .41] | -.44<br>[-1.24, .37] |
| Treatment, $\gamma_{01}$ | -.98<br>[-3.15, 1.20] | -1.11<br>[-3.39, 1.17] |
| Online, $\gamma_{02}$ | 3.83<br>[1.40, 6.25] | 3.77<br>[1.18, 6.36] |
| | | |
| Random Effects | | |
| Intercept, $u_{00}$ | 3.37<br>[1.00, 5.73] | 3.81<br>[.96, 6.67] |
| | | |
| Residual | 77.21<br>[64.74, 89.69] | 72.98<br>[60.13, 85.82] |
| | $N_{\text{school}} = 17$, $N_{\text{student}} = 1775$ | $N_{\text{school}} = 17$, $N_{\text{student}} = 1775$ |

Note. Model estimated in Mplus 7.11, restricted maximum likelihood estimator; all regression coefficients are unstandardised estimates

Table 15 presents the bootstrap estimates for the treatment effect from *b* = 1000 runs, weighted and unweighted as described above. As the bootstrap p-values of the table show, none of the effects of treatment reached the pre-specified significance level, and certainly not all of them together as the hypothesis suggested. In contrast to the normal theory-based confidence intervals in tables 14 and 15 above, the bootstrap confidence intervals indicate a potential for a negative effect of the intervention on the secondary outcomes, with three of the four bootstrapped confidence intervals for the effect sizes not including zero and being entirely in the range below zero.

Table 15. Bootstrap results for the fixed effect estimate of treatment for the hypothesis of potential transfer effects on the domains of Maths and English as a secondary outcome

| | Maths, Unweighted Analysis | Maths, Weighted Analysis | English, Unweighted Analysis | English, Weighted Analysis |
|---|---|---|---|---|
| Treatment Mean | -.977 | -1.153 | -1.657 | -1.501 |
| Treatment Median | -.977 | -1.137 | -1.648 | -1.492 |
| Lower 5%-bound | -1.837 | -2.000 | -2.476 | -2.434 |
| bootstrap p-value (percentage of coefficients > 0) | .038 | .009 | 0 | .002 |
| Mean effect size<br>[95%-Confidence Interval] | -.109<br>[-.227, .007] | -.132<br>[-.251, -.019] | -.150<br>[-.238, -.059] | -.134<br>[-.232, -.037] |

*Conclusion*

Based on the analyses for the secondary outcome (tables 9 + 10) it can clearly be said that no positive transfer effects of the program on English or Maths attainment occurred. This result also held, when the confidence intervals were bootstrapped and whether the analyses were weighted for missing data or not. The results rather pointed at a potential negative transfer effect, with students receiving LTSS receiving slightly lower scores than those in the control condition (effect sizes see table 11).

While the negative finding of positive transfer effects can be seen as quite robust and in line with the results obtained for the primary outcome, three caveats must be taken into account when gauging the potential for negative transfer effects:

(a) The mean effect sizes for both variables are to be considered small, and even the full study was not geared at showing effects of this size.

(b) The statistical analysis plan was geared at identifying positive effects, i.e. a single-sided testing strategy was originally proposed. Interpreting the results obtained here as evidence for a negative effect is therefore not entirely statistically correct.

**(c)** Although the test administration (English/Maths) was also randomly allocated across conditions and regions, the sample sizes within each condition are much smaller which makes it possible that the administration was confounded with other factors, which is impossible to test for with 17 / 20 schools.

### TERTIARY OUTCOME

To evaluate the degree to which the three instruments used in this study assess different aspects of academic attainment, the original plan was to analyse all three outcome measures together with confirmatory categorical data factor models[26]. This approach would have allowed treating the test data as categorical indicators of one or more latent attainment traits. Because of the change in the assessment schedule (see details in the protocol section) no schools received all three tests, therefore the factors "school" and "Maths vs. English" test are confounded and it is not possible to run this analysis as a clustered analysis anymore. Because of these drawbacks and based on feedback from the EEF this analysis was dropped.

[26] McDonald, R.P. (1999). Test theory: A unified treatment. Mahwah, NJ: LEA.

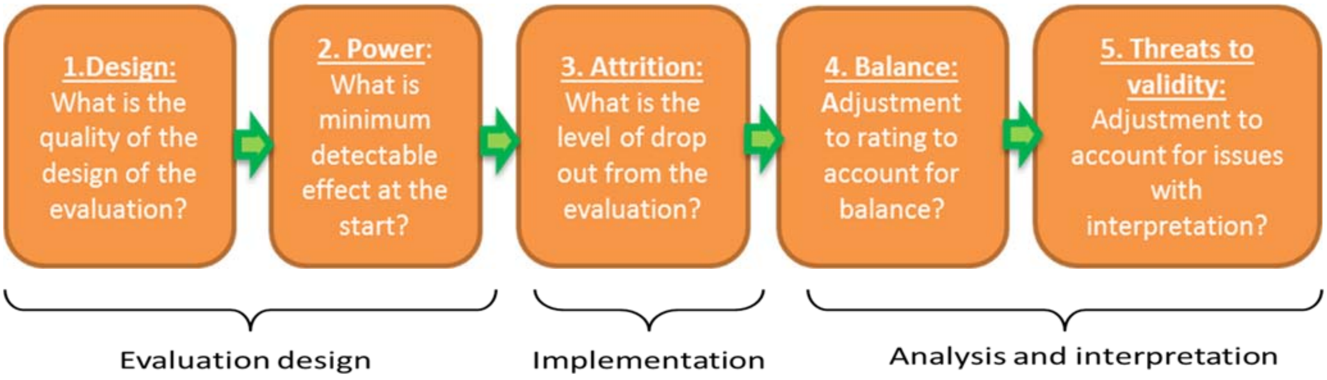
# Appendix J: Cost rating

Cost ratings are based on the approximate cost per pupil per year of implementing the intervention over three years. Cost ratings are awarded using the following criteria.

| Cost | Description |
|---|---|
| **£** | *Very low:* less than £80 per pupil per year. |
| **£ £** | *Low:* up to about £200 per pupil per year. |
| **£ £ £** | *Moderate:* up to about £700 per pupil per year. |
| **£ £ £ £** | *High:* up to £1,200 per pupil per year. |
| **£ £ £ £ £** | *Very high:* over £1,200 per pupil per year. |

# Appendix K: Padlock Rating

**7th July 2016 Complete by Elena Rosa Brown**


1.Design: What is the quality of the design of the evaluation?
2. Power: What is minimum detectable effect at the start?
3. Attrition: What is the level of drop out from the evaluation?
4. Balance: Adjustment to rating to account for balance?
5. Threats to validity: Adjustment to account for issues with interpretation?
Evaluation design
Implementation
Analysis and interpretation

| Rating | 1. Design | 2. Power (MDES) | 3. Attrition | 4. Balance | 5. Threats to validity |
|---|---|---|---|---|---|
| 5 🔒 | Fair and clear experimental design (RCT) | < 0.2 | < 10% | Well-balanced on observables | No threats to validity |
| 4 🔒 | Fair and clear experimental design (RCT, RDD) | < 0.3 | < 20% | | |
| 3 🔒 | Well-matched comparison (quasi-experiment) | < 0.4 | < 30% | | |
| 2 🔒 | Matched comparison (quasi-experiment) | < 0.5 | < 40% | | |
| 1 🔒 | Comparison group with poor or no matching | < 0.6 | < 50% | ↓ | ↓ |
| 0 🔒 | No comparator | > 0.6 | > 50% | Imbalanced on observables | Significant threats |

The final security rating for this trial is 3 🔒.

Education
Endowment
Foundation

The Education Endowment Foundation
9th Floor, Millbank Tower
21–24 Millbank
London
SW1P 4QP
www.educationendowmentfoundation.org.uk