Statistical Analysis Plan for the evaluation of Learning Counterintuitive Concepts

NFER

INTERVENTION	**Learning Counterintuitive Concepts (UNLOCKE)**
DEVELOPER	Birkbeck, University of London
EVALUATOR	National Foundation for Educational Research (NFER)
TRIAL REGISTRATION NUMBER	ISRCTN20284041
TRIAL STATISTICIAN	Stephen McNamara & Palak Roy
TRIAL CHIEF INVESTIGATOR	Simon Rutt
SAP AUTHOR	Stephen McNamara & Palak Roy
SAP VERSION	1
SAP VERSION DATE	20.04.2018

Protocol and SAP changes

No changes since the updated protocol was published.

Table of contents

Introduction

The Education Endowment Foundation (EEF) and the Wellcome Trust have commissioned the Birkbeck College to develop and deliver the counterintuitive concepts learning intervention in collaboration with the UCL Institute of Education.

The intervention *'Stop and Think'* is a computer-based programme and is designed to train pupils to use their ability of "interference control" to inhibit prior contradictory knowledge and misconceptions to acquire and use new knowledge successfully. Such an ability is required when learning new concepts in science and maths. The programme has the main aim of improving learner's ability to solve counterintuitive problems via training that will allow the individual to inhibit their initial response and instead, give a more delayed and reflective answer to ultimately improve learners' educational outcomes. It seeks to achieve these aims with Year 3 and Year 5 pupils receiving three 15-minute sessions a week at the start of a maths or a science lesson[1], where they use a teacher-led computer-based learning activity to practice counterintuitive problem-solving. In the game, a child-friendly character asks the player and other characters to solve problems, providing prompts and suggestions. Exercises will relate to specific maths and science content.

The project runs from January 2016 to December 2018 and is divided into two phases. The first developmental phase ran from January 2016 to July 2017. Aims of this phase were to develop the computer programme, to pilot it in eight[2] schools and finalise the mode of intervention delivery for the next phase. The aim of the second phase was to implement the intervention in 100 Primary schools and evaluate the impact of the intervention on a range of pupil outcomes. The SAP refers to this external evaluation by NFER using a cluster randomised controlled trial (RCT) design.

Study design

The population for this trial is all state-funded primary schools with at least one Year 3 class and one Year 5 class. The target was to recruit 100 primary schools predominantly but not exclusively, with above-average proportions of pupils receiving free school meals (FSM). In addition to this, the intention was to recruit 50 schools with one-form entry and 50 schools with more than one-form entry. All Year 3 and Year 5 classes from participating schools were required to take part in the trial. The evaluation is a cluster trial focused on Year 3 and Year 5 pupils. Year groups within schools were randomly allocated to one of the three groups making

[1] The teacher guides suggest that it is delivered at the start of either a maths or a science lesson. Although the theory of change suggests it should take place within a maths or a science lesson,it does not specify when it is delivered. We will explore this via the implementation and process evaluation.
[2] Note that only five of the eight schools continued with the pilot. A further three schools did complete a 'dress rehearsal' of the 10 week programme at class level prior to the intervention beginning.

it a three-armed cluster trial. The three groups are intervention, control and control plus. Overall, this meant that all schools would have at least one intervention group and either of the two control groups resulting in an unbalanced design with a ratio of 2:1:1 for the intervention versus the control or the control plus. The three trial arms can be described below:

1. Counterintuitive concepts in mathematics/ science lessons through a computer-based learning activity called '*Stop and Think*' also referred to as the 'intervention group'. This is intended to be run three times per week for 15 minutes at the start of maths or science lessons (the first 15 minutes of the lesson) and would run for 10 weeks.
2. 'Business-as-usual' control is referred to as the 'control group' where the normal classroom practice is continued.
3. Social skills control in PSHE lessons through a computer-based learning activity called '*See+*' also referred to as the 'control plus group'. This is a computer programme which captures the content of the age-appropriate PSHE and SEAL curricula. During this activity, children observe and reflect upon social interactions and engage in social-emotional learning through a series of computerised animated stories with virtual characters engaging in social scenarios. These sessions are intended to be run three times a week for 15-minutes and last 10 weeks. These can be delivered at the start of any lesson other than maths or science. The purpose of having a control plus group is to assess the impact of the computer programme (that is not the intervention).

The 10-week intervention will be class-based and teacher-led, as will be the control plus programme. This is an outcome of the pilot and an update to the original trial protocol which stated that either of the delivery methods were still possible.. The development and pilot phase of the study found this 1-to-1 delivery to be unfeasible for many schools resulting in these changes being made to the mode of delivery.

Birkbeck College was responsible for the recruitment for this trial which started in January 2017. Schools signed up to the trial via signing a memorandum of understanding (MoU). Once the school signed up to the trial, Birkbeck College collected administrative pupil data (pupil names, date of birth and UPNs). They supplied us with this data in order for us to access pupil characteristics such as FSM and prior attainment from the National Pupil Database (NPD)[3].

The primary research question is: does the use of the 'Stop and Think' intervention impact on learners' mathematics and science achievement? They will be measured by administering the progress test in Maths[4] (PTM) and the progress test in Science[5] (PTS) produced by GL Assessment. No baseline testing is required for this study since NPD is being used to access pupil prior attainment. Outcome measurements will take place between February and April 2018.

Randomisation

Whilst a target of 100 primary schools was in the original design, Birkbeck College was able to recruit 97 primary schools by signing the MoU. Of these, five schools were not eligible to take part in the trial as there was only one year group (either Year 3 or Year 5 or a mixed class across both the year groups), two schools didn't submit their administrative pupil data and one

[3] Foundation Stage Profile (FSP) data will be uiused as a prior attainment measure
[4] https://www.gl-assessment.co.uk/media/1346/ptm-technical-information.pdf
[5] https://www.gl-assessment.co.uk/media/1872/pts-technical_information.pdf

school withdrew participation prior to randomisation. As a result, 89 schools (178 year groups) were randomised.

Randomisation was carried out by a statistician at NFER using SPSS software with a full syntax trail. This was done in two waves during October 2017 to accommodate staggered time for installing the computer programme(s) and training both of which were undertaken by the research assistants from Birkbeck College. The software installation took place in all schools as at least one class in each school would be allocated to the intervention. No baseline testing was required since the Foundation Stage Profile (FSP) is being used as a prior attainment measure for analysis. This means these data were collected prior to randomisation.

Each eligible school had at least one Year 3 class and one Year 5 class. The unit of randomisation was the year group within a school. The randomisation was stratified by the number of form entry as it was important to ensure that the number of intervention classes was balanced with a similar number of classes in control and control plus groups together. If the randomisation had not been stratified by form entry, control or control plus may have become disproportionately represented in three-form entry schools[6], e.g., with likely higher numbers of pupils.

Schools were randomised as one of four possible set-ups:

- One-form entry, both years
- Two-form entry, both years
- Three-form entry, both years
- All other form entry (i.e. 1x Year 3 and 2x Year 5, 2x Year 3 and 3x Year 5, 4x Year 3 and 4x Year 5)

Table 1 presents the number of schools and year groups randomised to each trial arm. As explained earlier, for every school, one year group was assigned to the intervention group and another year group was assigned to either of the two control groups. For example, for the first wave, in 28 schools Year 3 was assigned to the intervention group, in 15 schools Year 3 was assigned to the control group and in 16 schools Year 3 was assigned to the Control Plus group.

Table 1. Number and proportion of schools and year groups randomised

	Trial arms	Year 3	Year 5	Total Year Groups
Wave 1 (59 Schools, 118 Year Groups))	Intervention	28	31	59 (50%)
	Control	15	13	28 (24%)
	Control Plus	16	15	31 (26%)
Wave 2 (30 Schools, 60 Year Groups)	Intervention	14	16	30 (50%)
	Control	7	7	14 (23%)
	Control Plus	9	7	16 (27%)

[6] In recruiting schools some were a variety of form structures that were different from the one and two class entry system as originally specified, and this needed to be accounted for in the randomisation process.

Total (89 Schools, 178 Year Groups)	**Intervention**	42	47	89 (50%)
	Control	22	20	42 (24%)
	Control Plus	25	22	47 (27%)

Overall, there is an imbalance in the group allocation for year groups in control and control plus. This occurred as a result of not correcting the group imbalance that arose in the first wave. We deliberately adopted this approach to ensure that bias is not introduced in case the schools in the second wave are systematically different from the schools in the first wave. By not correcting this imbalance, we are not allowing more 'control groups' than 'control plus groups' in the second wave schools.

Shortly after randomisation, two schools withdrew[7] without the knowledge of group allocation. Since this presented no source of bias, they were removed from the trial. The final number of schools retained in the trial was 87.

Calculation of sample size[8]

Sample size from the protocol

At protocol[9], power calculations used the following two assumptions which were obtained from EEF's paper on pre-test effects (EEF, 2013). The initial design believed KS1 could be used as a covariate but due to changes in how KS1 is measured and reported it was decided to use a measure that would be consistent for both of our year groups. FSP[10] would therefore be used as a covariate in the analysis with GL Assessment's Progress in Maths and Progress in Science being used to measure primary outcomes. The correlation between FSP and these assessments in Year 3 and Year 5 is assumed to be 0.65. The rationale for selecting a correlation of this size was based on a paper produced by the Fisher Family Trust (FST)[11]. The intra-class correlation (ICC) is assumed to be 0.126 (EEF, 2013), as this was the ICC identified for maths total score at KS2 Any values below this will increase the design's power for the given effect sizes. These figures were used in the calculation of optimum sample sizes for desired levels of power. These assumptions allowed for the following comparisons:

Example Table 3: Minimum detectable effect size at different stages[12]

Stage	N [schools/pupils] (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)

[7] All the reasons for school dropouts will be identified in the final report.
[8] We use our own excel formula to calculate the MDES.
[9] A sample of 100 schools had been identified at the awarding of the initial grant, prior to the evaluators being in place. The subsequent design using 100 schools results in a relatively low MDES and school attrition would result in a higher MDES. Calculations suggest that an effect size of 0.15 could still be detected with 30% attrition.
[10] The protocol will be amended to reflect this change
[11] http://csapps.norfolk.gov.uk/csshared/ecourier2/fileoutput.asp?id=11608
[12] This table identifies the MDES for the main analysis between the intervention and comparison groups. Other MDES calculations are included in the body of the text. The table also identifies the MDES for one of primary outcomes, although assumptions and numbers are the same for both measures.

Protocol	100/4050; (100/2025,100/2025)	0.65	0.126	School blocking	80%	0.05	0.125
Randomisation	87/3265;(87/1612, 87/1653	0.65	0.126	School blocking	80%	0.05	0.135
Analysis (i.e. available pre- and post-test)							

- n (intervention) =100 schools and 150 classes; n (control and control plus) =100 schools and 150 classes represented the comparison between intervention classes and both control and control plus classes grouped together and assumed an average cluster size of 27 (average cohort size for eligible primary schools class in England). Power calculations were based on half of these pupils taking a maths test and the other half taking a science test. Calculations were based on an effect size for either of these tests. Both assessments are therefore powered to 80%.

- n (intervention) =100 schools and 150 classes; n (control) =50 schools and 75 classes represents the comparison between the intervention classes and the control plus group. This again assumes an average cluster size of 27 (average cohort size for eligible primary schools in England).

- n (control) =50 schools and 75 classes; n (control plus) =50 schools and 75 classes represents the comparison between the control and control plus groups. This assumes an average cluster size of 27 for the size of each class.

The main trial was well powered with a minimum detectable effect size (MDES, at 80% power) of less than 0.2 for all three types of analyses. The intervention and control/control plus comparison (main analysis) had MDES of around 0.125. The MDES for intervention and control plus analysis is 0.152 and the MDES for control and control plus is 0.176, both at 80% power.

Assuming that there were 22.5% pupils who are eligible for FSM at any time during the past six years, the MDES for FSM only analysis would be 0.17 at 80% power.

MDES after recruitment

Following the recruitment and randomisation, 87 schools have been retained in the trial. One more school withdrew from the trial. As they knew the year group allocations, it would be important to keep the school in the analysis. However, the school is not willing to take part in the primary outcomes tests. Therefore, the total number of schools would be 86. Sample size calculations have been re-run and the revised MDES for each of the analyses are 0.135, 0.17 and 0.19 respectively. Revised MDES for FSM only analysis has increased slightly to 0.19 with 80% power.

Follow-up

As mentioned earlier, 89 schools were recruited and randomised. Of these, two schools dropped out of the trial without the knowledge of group allocation. Following randomisation,

another school dropped out of the trial. As this school was aware of the group allocation, this would be considered a biased drop-out. As a result, 87 schools were followed up for the primary outcome testing.

Birkbeck College collected administrative pupil data in order for schools to be included in the randomisation. Before passing this data to Birkbeck College, schools administered parental opt-out so that schools did notsupply pupil data in the case where they had received a parental opt-out. We received administrative pupil data for 6,530 pupils across 87 schools from Birkbeck College. Of these, schools had received parental opt-out for further fourteen pupils across ten schools. These pupils will not be included in the pupil list sent to NPD for data matching and therefore will have missing prior attainment and other background characteristics.

Two schools had their year groups incorrectly assigned to the treatment conditions as a result of mixed communication from Birkbeck College. These schools are being treated and will be analysed as randomised to fulfil an intention to treat analysis. A further two schools, one from each wave, were randomised under incorrect form set up[13]. This does not affect the implementation of the intervention but causes a small imbalance across the randomised groups.

Outcome measures

Primary outcome

There are two primary outcome measures for this trial[14]. They will be measured by administering the PTM and the PTS produced by GL Assessment. NFER will manage the test administration by sending the test administrators to schools in February and March2018. This is to ensure that the tests are administered blind to group allocation and would reduce burden placed on schools. As there are two separate year groups, it is necessary to administer age-appropriate tests. Year 3 pupils will take PTM8 and PTS8, Year 5 pupils will take PTM10 and PTS10. The power calculations were based on each pupil taking only one subject test. Within each class, half the students were randomised to take a maths test and the other half will take a science test. Randomisation was undertaken by an NFER statistician. This was a simple randomisation allocating equal number of pupils within a class to maths or science test. Schools were sent the pupil allocation to maths or science test one week prior to testing.

Raw total scores from the PTM and PTS will be used as the primary outcome measures. These outcome measures will be analysed and reported separately. Maximum possible score for PTM8 is 55, PTM10 is 65, PTS8 is 40 and PTS10 is 50. On all the assessments, higher score indicates higher attainment.

As Year 3 and Year 5 pupils will take different assessments, it will be necessary to analyse outcomes from these assessments separately. For example, for maths, outcomes from PTM8 and PTM10 will be analysed in separate models. Effect sizes from these models will be combined to determine an overall impact of the intervention on pupils' attainment in maths. This combined effect size will constitute the primary outcome measure in maths. Similar

[13] These two schools were randomised as mixed-form entry schools where they are, in fact, one-form entry schools with one class in each year group.
[14] Note that there were two primary outcomes options in the original protocol. After the development phase was complete, it was decided that both the outcomes will be retained as the primary outcome measures for the trial as suggested in the original protocol.

analysis will also be undertaken for science, See analysis section for details on the analytical methods.

Secondary outcomes

The secondary outcome measure for the trial will be assessed using a new Stroop-like measure of inhibitory function development. This assessment is drawn from Wright *et al.* (2003). The pupils will work through five sheets (1 practice, 2 congruent condition and 2 mixed conditions), each to be completed as well as possible within 10 seconds. In the congruent condition the animal head matches with the animal body. In the mixed condition half the animals have heads that match their bodies and the other half will have heads that do not match their bodies. It is the latter that will enable us to assess their absolute performance in the mixed conditions. The raw total score from the mixed sheets will be used for the analysis. This score ranges from 0-30, where a higher score indicates better attainment. The raw total score from the congruent sheets will be included in the model to control for cognitive skills not related to inhibitory control. This test will be administered by the research assistant appointed by Birkbeck College[15] and will take place after the schools have completed the GL assessment tests.

Analysis

The trial analysis will follow the latest EEF Analysis Guidance[16]. Uncertainty around the effect sizes will be presented as per the framework and as described in the effect size section below.

Primary intention-to-treat (ITT) analysis

As previously mentioned, there will be separate analysis for each primary outcome measure- one for maths and the other for science. The overall impact of the intervention on pupils' attainment in a given subject will be determined by combining the effect sizes from the two-year group models. For maths, outcomes from PTM8 (Year 3) and PTM10 (Year 5) will be analysed in two separate models. This combined effect size will constitute the primary outcome measure in maths. Similar models will be run and combined to determine an overall effect size for the primary outcome measure in science. Model details and the calculation of effect sizes are described below.

The primary outcomes analyses will be 'intention-to-treat' and will be conducted at pupil level, comparing average pupil maths or science scores in the intervention group with average scores in the control or control plus (combined) group. As the pupil level data will be clustered within classes that are clustered within year groups and schools, the hierarchy of the data will need to be acknowledged in the models. Each model will be run at year group so year group will not be included as one of the levels. Therefore, multilevel linear regression models with three levels (school, classes and pupils) will be used to analyse the impact of the intervention on pupil outcomes.

As per the protocol, the original plan was to use Key Stage 1 (KS1) assessment data as pupil prior attainment. Since the new assessment and reporting arrangements were introduced for

[15] Birkbeck College will ensure that the research assistants are blind to group allocation of the year groups they are administering the assessments with.

[16]https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Protocol_or_SAP/EEF_statistical_analysis_guidance_2018.pdf

KS1 from 2015-16, this affected the KS1 attainment data in terms of what was collected from schools and made available in the NPD. This means, the prior attainment measures of KS1 would not be consistent for the year groups in the trial. Year 5 pupils took the KS1 tests under the old system in 2014-15 and Year 3 pupils took the KS1 tests under the new system in 2016-17. In the absence of any comparable measures of KS1, we will use the closest possible measures to control for pupil prior attainment that is consistent across the year groups in the trial. We will use average FSP point score as measured by combining all 17 early learning goals. These variables are available on the NPD with a value range of 1-3 where higher scores reflect higher attainment for a given goal.

Maths outcome

In Year 3 maths model, the dependent variable will be PTM8 raw total score with the following covariates:

- An indicator of whether the pupil was in the intervention group (reference category = combined control group that consists of both control groups)
- stratification variable used at randomisation to indicate whether the school is a two-form entry, three-form entry or mixed-form entry school (reference category=one-form entry school)
- Foundation Stage Profile score will be used as a prior attainment measure

In Year 5 maths model, the dependent variable will be PTM10 raw total score with the following covariates:

- An indicator of whether the pupil was in the intervention group (reference category = combined control group that consists of both the control groups)
- stratification variable used at randomisation to indicate whether the school is a two-form entry, three-form entry or mixed-form entry school (reference category=one-form entry school)
- Foundation Stage Profile score will be used as a prior attainment measure

The overall effect for the maths outcome will be an amalgamation of the effects of Year 3 and Year 5 models. These will be combined according to the method described on page 227 of Borenstein et al. (2009). This allows the combination of non-independent effects for the same trial. Since the outcome measures of Year 3 and Year 5 maths include the same schools, it is important not to assume that these outcomes are separate and providing independent information. Year 3 mean scores on maths may be correlated with Year 5 mean score on maths and thus the effect sizes for both the year groups are interdependent. The combined variance calculation will take this correlation into consideration to properly estimate the precision of the overall effect. We will use the following formula to amalgamate the two effects sizes where Y_{m3} and Y_{m5} are the effects sizes from the Year 3 and Year 5 models respectively and Y_c is the combined effect size.

$$Y_c = \frac{1}{2}(Y_{m3} + Y_{m5})$$

We will use the following formula to calculate the variance for the combined effect size where V_{m3} and V_{m5} are the variance from the Year 3 and Year 5 maths models respectively,

Vc is the variance for the combined effect size and r is the correlation coefficient that describes the extent to which Year 3 maths score and Year 5 maths scores co-vary.

$$V_c = \frac{1}{4}\left(V_{m3} + V_{m5} + 2\mathrm{r}\sqrt{V_{m3}}\sqrt{V_{m5}}\right)$$

Science outcome

Similar to maths, two models will be run for the science outcomes.

In Year 3 science model, the dependent variable will be PTS8 raw total score with the following covariates:

- An indicator of whether the pupil was in the intervention group (reference category = combined control group that consists of both control groups)
- stratification variable used at randomisation to indicate whether the school is a two-form entry, three-form entry or mixed-form entry school (reference category=one-form entry school)
- Foundation Stage Profile score will be used as a prior attainment measure

In Year 5 science model, the dependent variable will be PTS10 raw total score with the following covariates:

- An indicator of whether the pupil was in the intervention group (reference category = combined control group that consists of both control groups)
- stratification variable used at randomisation to indicate whether the school is a two-form entry, three-form entry or mixed-form entry school (reference category=one-form entry school)
- Foundation Stage Profile score will be used as a prior attainment measure

Similar to the maths outcome, the overall effect for the science outcome will be an amalgamation of the effects of Year 3 and Year 5 science models.

All the data manipulation will take place in SPSS and the models will be run in R.

Imbalance at baseline for analysed groups

We expect no systematic bias to have arisen from randomisation. However, it is important to explore whether pupils assigned to treatment conditions differ based on background characteristics. Imbalance in the group allocation (as assigned at randomisation) will be explored in regards to background characteristics such as pupil FSM eligibility and prior attainment. We will use multilevel modelling to examine imbalance for prior attainment where these models will have similar structure to the ITT models with three levels schools, classes and pupils. There will be two separate models- one for each year group. Prior attainment will be regressed on whether the pupil belonged to the intervention or control group.

Missing data

We will assess whether missing data at the randomisation level of year groups comprise more than a threshold amount of 5% of the total data. If this is found to be larger than 5% from either

of the two randomisation groups (intervention and combined control groups), we will carry out further analysis. In particular, a logistic multilevel model of whether or not an individual is missing regressed on the covariates of the main model. This will help determine the extent of bias.

Under the 'missing at random' assumption, we would expect a completers analysis to be unbiased. If the extent of dropout was unequal between the randomised groups, the 'missing not at random' assumption is likely to hold and we will conduct sensitivity analyses. This will be done by initially running multilevel multiple imputation. Following analyses undertaken on other EEF funded evaluations, we would propose a methodology that includes all the variables included in the primary analysis plus other variables available from the NPD to run models that identify the significant variables associated with missingness. These significant variables would then be used for a multiple imputation process using the mice package in R. The number of datasets is dependent on the amount of missing data but a minimum would be five datasets, with a minimum of ten iterations. These iterations are necessary as with only one dataset, the parameter estimates have more sampling variability. Multiple iterations also help in generating the estimates of the standard errors to accurately reflect the uncertainty about the missing values (Allison, 2012). The model would then be extended using a weighting approach according to Carpenter et al. (2007). Missing data analysis will only be possible in cases where we have pupil administrative data and a subsequent match with the NPD.

Non-compliance with intervention

It is likely that not all sessions will proceed exactly as planned. Since the intervention is computer-based, it will be possible to extract an exact number of sessions performed by each teacher or completed by each class. The main analysis will, therefore, be followed by a CACE analysis (Complier Average Causal Effect) in order to assess the effect of non-compliance on outcome measures where data from the computer system will be used to determine the extent of each class's involvement. The information on a number of completed sessions will be collected by Birkbeck College as agreed with NFER. This will determine the compliance or engagement level of each class. Although the compliance measures are at class level, the unit of analysis will be pupils. Following table presents how the compliance will be measured. These measures will be included in the analysis as ordinal variables.

Level of compliance	Description (in numbers of completed sessions)
None	0 sessions
Low	1 to 10 sessions
Medium	11 to 20 sessions
High	21 to 30 sessions

Schools may potentially have unobserved characteristics that have an influence on both the compliance with the intervention and academic attainment. Therefore, a two-stage least squares model will be used to calculate the CACE estimate (Angrist and Imbens, 1995). The first stage of the model will be compliance regressed on all covariates that are used in the main primary outcome model and in addition, will include, as an instrumental variable, a binary variable that indicates a pupil's pre-intervention treatment allocation. The second stage of the model will regress the primary outcomes on the covariates used in the main models and will also include a covariate representing the pupil's estimated level of compliance from the first

stage of the model and an interaction term between the estimated compliance and the pupil's pre-intervention treatment allocation. The coefficient of the interaction term is the CACE estimate of the compliance effect. In the event that there are no confounding factors affecting compliance and attainment the CACE estimate will be equal to the intention-to-treat estimate. We will use the R package ivpack to perform the CACE analysis on the primary outcomes only.

Secondary outcome analyses

The outcome of the animal-Stroop task will be analysed via multilevel linear regression models. Analyses will be performed at pupil level, in a three-level hierarchy to account for clustering within classes and schools. Two separate models will be run- one for each year group. The dependent variable in these models will be the raw total score from the Stroop task regressed on the following covariates:

- An indicator of whether the pupil was in the intervention group (reference category = combined control group that consists of both control groups)
- stratification variable used at randomisation to indicate whether the school is a two-form entry, three-form entry or mixed-form entry school (reference category=one-form entry school)
- Raw total score in the congruent sheets as a control for non-inhibitory control cognitive skills

The combined effect size from the two year group models will determine the overall impact of the intervention on this outcome of inhibition control. As discussed earlier, these will be combined according to the method described in Borenstein et al. (2009).

Additional analyses

Two additional analyse are planned. These will be performed on a subset of pupils in two of the three arms of the trial. The first analysis will look at differences between the intervention and the control plus group and the second analysis will look at differences between the control plus group and the business as usual control group.

These models will be similar to those discussed in the primary analyses where two separate effect sizes will be reported, one for each subject and both the effects will need to reach the statistical significance. The year group models will be run separately and the combined effect size will constitute the overall effect of the intervention on given subject outcome measure.

Subgroup analyses

As specified in the protocol, sub-group analyses on the primary outcomes will take place to explore the impact of individual characteristics. As per the EEF guidance, there will be an interaction model of whether a pupil has ever received free school meals (as measured by EVERFSM_6 variable from the Autumn School census 2017-18). This will be done using models identical to the primary outcomes models but including EVERFSM_6 and EVERFSM_6 interacted with the intervention indicator as covariates. Analyses shall proceed as per the original primary outcomes modelling. A separate analysis of FSM only pupils will

also be carried out as per the EEF analysis guidance. These models will be similar to the main models of overall effect but will only include pupils who were eligible for FSM as measured by EVERFSM_6 variable.

Another interaction models will also be run to detect differential impact based on pupil gender. An interaction term will be added to the main models along with gender variable- intervention interacted with pupil gender. Analyses shall proceed as per the original primary outcomes modelling.

Age will not be included in these models as separate models will be run for each year group. This is an amendment from the original protocol and this will be included in a protocol amendment to reflect this change.

Data manipulation will be carried out in SPSS while the multilevel models will be run in R package nlme and imputation macros available from missingdata.org.uk.

Effect size calculation

For the primary outcomes, two effect sizes will be produced. The effect size for maths will be produced by combining the effect size observed in the Year 3 maths and the Year 5 maths models. Similarly, the effect size for science will be produced by combining the effect size of the Year 3 science and Year 5 science models (Borenstein et al., 2009).

The numerator for each individual model effect size calculation will be the coefficient of the intervention group from the multilevel model. All effect sizes will be calculated using total variance from the multilevel models, without covariates, as the denominator i.e. equivalent to Hedges' g. Confidence intervals for each effect size will be derived by multiplying the standard error of the intervention group model coefficient by 1.96. These will be converted to effect size confidence intervals using the same formula as the effect size itself.

Report tables

All the tables will be structured according to the EEF trial report template[17].

References

Allison, P. (2012). 'Why you probably need more imputations than you think', *Statistical Horizons*, 9 November [online]. Available: https://statisticalhorizons.com/more-imputations [9 March, 2018].

Borenstein M, Hedges LV, Higgins JPT and Rothstein HR. (2009). *Introduction to Meta-Analysis*. Chichester: John Wiley & Sons, Ltd [online]. Available: http://onlinelibrary.wiley.com/book/10.1002/9780470743386 [31 January 2018].

Carpenter, J.R., Kenward, M.G. and White, I.R. (2007). 'Sensitivity analysis after multiple imputation under missing at random: a weighting approach', *Statistical Methods in Medical Research*, **16**, (3), 259-275.

Education Endowment Foundation (2013). *Pre-testing in EEF Evaluations*. London: EEF [online]. Available:

[17] https://educationendowmentfoundation.org.uk/evaluation/resources-centre/writing-a-research-report/

https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Protocol/Pre-testing_paper.pdf [31 January 2018].

Wright, I., Waterman, M., Prescott, H. and Murdoch-Eaton, D. (2003). 'A new Stroop-like measure of inhibitory function development: typical developmental trends', *Journal of Child Psychology and Psychiatry*, **44**, (4), 561-575.