# Statistical Analysis Plan for Philosophy for Children

NFER

| | |
|---|---|
| INTERVENTION | **Philosophy for Children (P4C)** |
| DEVELOPER | SAPERE |
| EVALUATOR | National Foundation for Educational Research (NFER) |
| TRIAL REGISTRATION NUMBER | ISRCTN11118203 |
| TRIAL STATISTICIAN | Jack Worth |
| TRIAL CHIEF INVESTIGATOR | Dr Ben Styles |
| SAP AUTHOR | Palak Roy & Constance Rennie |
| SAP VERSION | 1 |
| SAP VERSION DATE | 06.10.2017 |
| EEF DATE OF APPROVAL | 06.10.2017 |
| DEVELOPER DATE OF APPROVAL | 23/10/17 |

## Protocol and SAP changes

The randomisation section of the protocol states that simple randomisation will be used (across two large blocks). The analysis section of the protocol contains a reference to including a ‘set of region dummy variables’ to account for stratified randomisation. The randomisation section is correct i.e. simple randomisation was used and no region dummies will be required in the analysis. Stratifying the randomisation by region was considered in the design phase, to aid intervention delivery across the country. However, simple randomisation is preferred for the analysis because fewer degrees of freedom are lost to control for the stratification. Simple randomisation was used once it was established that stratification by region was unnecessary for intervention delivery.

More detail on which social skills measures to analyse at follow-up was added to the analysis specification that was not included in the protocol.

# Introduction

Philosophy for Children (P4C) is an approach to teaching in which students participate in group dialogues focused on philosophical issues. Dialogues are prompted by a stimulus (for example, a story or a video) and are based around a concept such as 'truth', 'fairness' or 'bullying'. The aim of P4C is to help children become more willing and able to ask questions, construct arguments, and engage in reasoned discussion. P4C was originally developed by Professor Matthew Lipman in New Jersey, USA in 1970 with the establishment of the Institute for the Advancement of Philosophy for Children (IAPC).

The Society for the Advancement of Philosophical Enquiry and Reflection in Education (SAPERE), a non-profit society, promotes the use of P4C in UK schools along with developing teaching resources and providing teacher training courses. P4C is practised across all education age ranges. SAPERE's model of P4C differs in some ways from Lipman's original conception. In particular, there is no use of specially written philosophical novels. Materials recommended by SAPERE include stories, poems, scripts, short films, images, artefacts, and picture books. However, Lipman's central aim of creating a classroom 'community of enquiry' is retained along with the broad sequence of activities that constitute a P4C session.

P4C has been the subject of a number of studies since the 1980s, these have had various methodologies, but have consistently shown impacts on logical reasoning and reading. Gorard et al. (2015) represented the first large-scale evaluation of the impact of P4C on attainment in English schools. The study in 48 schools showed that P4C had a positive impact on Key Stage 2 attainment, with pupils using the approach making approximately two additional months' progress in reading and maths. The Gorard et al. (2015) trial was classified as an effectiveness trial, meaning that it sought to test whether the intervention can work at scale. However, because of the relatively small number of schools involved (48 schools), this study aims to obtain a more secure estimate of the impact of P4C on all children and particularly on children eligible for free school meals.

# Study design

### *Description of population including eligibility criteria*

Junior and primary schools that include pupils in year groups four, five and six will be considered for eligibility in the trial. Schools recorded in the 2015 annual school census data as having more than 25 per cent of their pupils that have ever been eligible for free school meals (EVERFSM-eligible) and have not previously implemented whole-school P4C will be eligible for the trial.

### *Description of trial design*

The impact evaluation will use a cluster-randomised design to identify the causal effect of the intervention on attainment. Schools will be randomly allocated to receive the intervention or business-as-usual control. Schools allocated to the intervention group will receive training and support over three years, while schools allocated to the control group will continue teaching as normal and will be asked not to use P4C materials during the first two years, and to abstain from using P4C materials with year 6 pupils in the third year of the trial[1].

[1] Some contamination of the control group in the third year is possible with this approach, which is why the primary outcome of the trial is attainment at the end of the second year of the trial.

## Sample size

The planned sample size for the trial was 200 junior and primary schools, with 75 randomised to receive the intervention and 125 to be part of a control group. The number of intervention schools was kept at no more than 75 to aid intervention delivery. An unbalanced design that includes more than 75 control schools improves the precision of the estimate of impact compared to a balanced design with 75 control schools (see 'calculation of sample size' section below). In total, 198 schools were recruited and randomised for this trial: 75 to intervention and 123 to control.

## Description of trial arms

Intervention schools will receive training and support over three years, with the intention of reaching SAPERE's Gold Award level of P4C practice by 2020. For each school, the programme will consist of the following elements:

- 2 days of P4C Foundation Training (Level 1) for up to 25 staff:  this equips teacher to start facilitating P4C enquiries with their students, and covers the basic principles of P4C practice, the standard enquiry model and provides an opportunity to experience a model enquiry;
- 1 day of P4C Tools for Thinking Together Training for up to 25 staff; this provides staff with additional facilitation techniques and practical guidance in encouraging stronger reasoning and conceptual thinking among students:
- 4 days of Advanced P4C Training (Level 2A and 2B) for 2 staff; Level 2A gives the school's P4C leaders advanced facilitation techniques so that they can support colleagues who are less advanced in their P4C practice; Level 2B gives the P4C leaders guidance in how to plan for the development of the school's P4C practice, how to link P4C into the broader curriculum and how to handle sensitive and controversial topics that may arise in an enquiry;
- 7 days of in-school P4C coaching and support;  the SAPERE trainer tailors the content of these days to the school's needs;  they may include demonstration, observation or co-teaching by the trainer, or planning with the P4C leader or remedial work with teachers who need extra assistance, or specialist advice on linking P4C to literacy, for example;
- 5 days of remote administration and planning support; these are for ad hoc support on the implementation of P4C and may include guidance on the Bronze, Silver and Gold award applications;
- Unlimited access to SAPERE's online P4C resources and practice guides;  these include a wide bank of suggested enquiry stimuli, a Getting Started Guide, a Moving On with P4C guide, a range of teaching materials and example enquiry plans and the Award framework which sets out a detailed progression for P4C practice across student, teacher and whole school dimensions;
- 2 reference copies of SAPERE's Level 1 and Level 2 handbooks;
- Application and assessment fees for SAPERE's Bronze, Silver and Gold awards.

The initial training will be delivered as INSET days with up to 25 teaching staff between March and October 2017. These teachers will introduce weekly P4C sessions for Year 4, 5 and 6 classes from September 2017 onwards.

Schools from the business-as-usual control group will continue teaching as normal and will be asked not to use P4C materials during the first two years, or for Year 6 pupils in the third year.

## Randomisation

Schools were randomised to the intervention or control group by simple randomisation. There were two randomisations of schools:  one in January 2017 and another in March 2017. The first randomisation was conducted to meet a commitment to a randomisation date made in the information materials initially sent to schools and the second randomisation was made to allow more schools to enter the trial and for the desired number of recruited schools to be met. The first randomisation divided 110 schools into intervention and control according to the expected 75:125 overall ratio: 41 schools were randomised to the intervention group and 69 were randomised to the control group.

Since there were 198 schools recruited, we adjusted the allocation ratio slightly in the second randomisation block in order to increase statistical power. We allocated a total of 75 schools to the intervention group and the 123 remaining to the control. This adjustment was carried out as planned in the protocol. The second randomisation block therefore included 88 schools: 34 were randomised to the intervention group and 54 were randomised to the control group. Randomisation results are reported in Table 1.

**Table 1: Number of schools randomised**

| | Block 1 | Block 2 | Total |
|---|---:|---:|---:|
| **Intervention group** | 41 | 34 | 75 |
| **Control group** | 69 | 54 | 123 |
| **Total** | 110 | 88 | 198 |

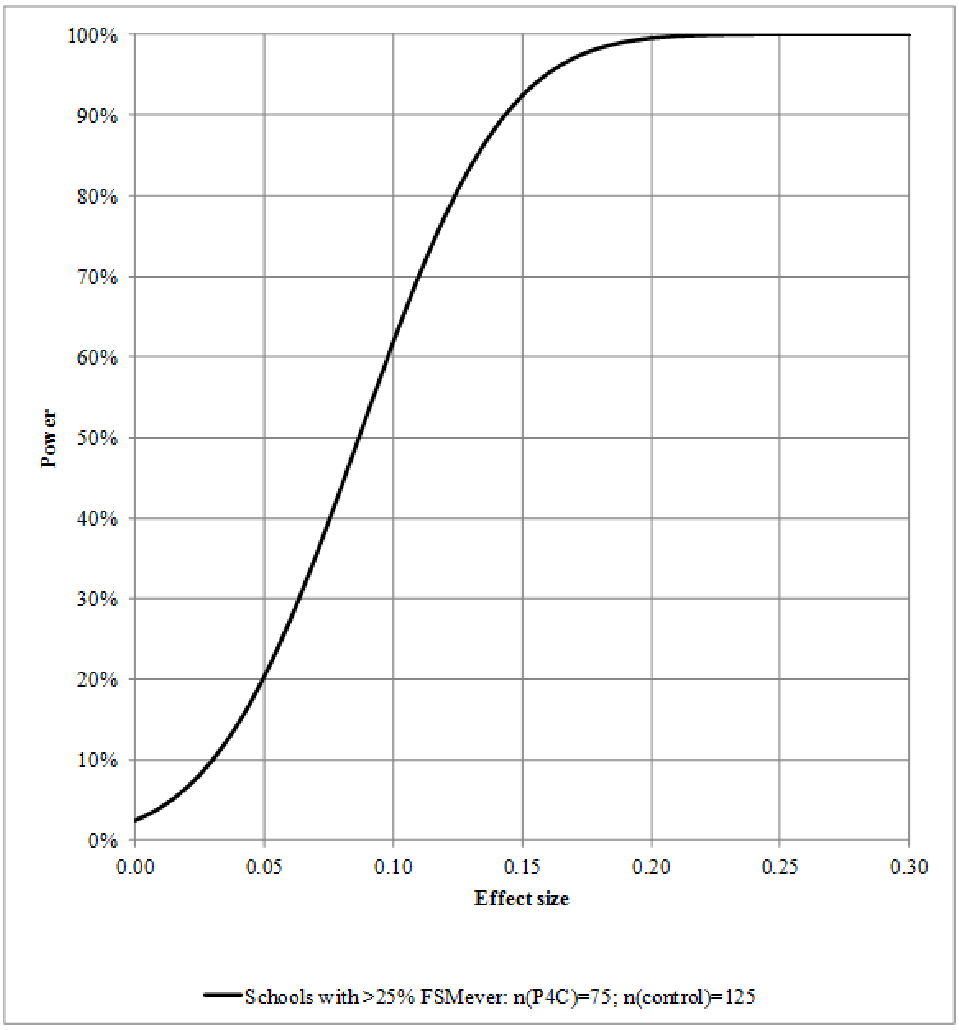
Randomisation was carried out by a statistician at NFER using SPSS software. The randomisation syntax will be published in the final evaluation report.

## Calculation of sample size

The required sample size was determined by the need to ensure the design can detect a reasonably small effect size (0.125) among EVERFSM-eligible pupils. The capacity of the developer to deliver training across a large number of schools across different areas of England during two school terms was also taken into consideration when deciding the ratio of intervention to control schools.

The aim of the evaluation was to recruit 200 eligible schools to participate, 75 schools to be randomly allocated to the intervention group, while 125 schools to be allocated to the control group. Given our assumptions about the number of EVERFSM-eligible pupils per eligible school, the intra-cluster correlation and the correlation between pre-test and post-test, the original design was powered to detect an effect size of 0.125 among EVERFSM-eligible pupils. Balancing the proportion of intervention and control schools (100 vs 100) would have given the design marginally higher power (83.0% compared to 80.5%), but this could have caused delivery issues for the developer. Balancing the sample size with 75 intervention schools and 75 control schools would have under-powered the trial design. Therefore, a sample size was chosen that optimised the delivery capacity without a huge impact on the statistical power. The power curve from the chosen design is shown in Figure 1.

**Figure 1: Power curve from the protocol**



As there were 198 schools recruited for the trial, 75 schools were assigned to the intervention and 123 schools were assigned to the control group (two less than originally planned). This reduction in the number of control schools affects the statistical power only slightly, reducing it from 80.5% to 80.3%.

## Follow-up

Follow-up data on pupil outcomes will come from two main sources. Data for the primary outcome (reading attainment) and one secondary outcome (mathematics attainment) will come from the National Pupil Database (NPD). Data for the social skills secondary outcome will come from a questionnaire administered by NFER.

Since the primary outcome data comes from the NPD, we can define the definitive list of eligible pupils from pupils enrolled in the participating schools in the school census. As the randomisation took place in January and March 2017, pupils on roll from all participating schools in the spring term 2016/17 school census will constitute the definitive pupil list.

School drop-out from the trial will not affect the primary outcome measure (the 2019 Key Stage 2 reading scaled score for EVERFSM pupils) as this comes from the National Pupil Database (NPD). We will be obtain NPD data for all pupils who were in years 3 and 4 at the participating schools in the spring term 2016/17. Follow-up data for the primary outcome and other attainment outcomes will come directly from the NPD and will be unaffected by any school drop-out from the trial. Indeed, the outcomes of pupils in the definitive list who move to non-participating schools can still be obtained from the NPD, providing they are in the English state-funded sector in year 6.

As the non-attainment secondary outcome measure will be collected from a social skills questionnaire in the summer term 2018/19 (the same cohort as the primary outcome measure), it might be affected by schools that drop out of the trial. In the event of a school wishing to withdraw their participation from the trial, as far as possible, we will try and collect this measure via the pupil questionnaire. The questionnaire will be administered by NFER's test administrators, which reduces the burden on schools and also ensures a good return rate. If a school is not willing for NFER's test administrators to administer the pupil questionnaire, we will not be able to collect this measure from the pupils at that school. This will constitute missing data at school level. Missing data at pupil level will constitute those pupils who were absent on the day of administration (at baseline or end-point), or on school roll in the spring term 2016/17 but moved to a different school prior to the summer term 2018/19. Detailed discussion on how the missing data will be handled is included in the analysis section.

## Outcome measures

### *Primary outcome*

The primary outcome measure of the impact evaluation will be the 2019 Key Stage 2 reading scaled score for EVERFSM pupils. These pupils were in year 4 in 2016/17 and will be in year 6 in 2018/19. Variable KS2_READSCORE will be used from the 2018/19 Key Stage 2 NPD. EVERFSM pupils will be identified by using the EVERFSM_6_P variable from the spring term school census of 2016/17 available from the NPD.

### *Secondary outcomes*

Secondary attainment outcomes will consist of Key Stage 2 maths scaled score for the same cohort as the primary outcome measure. Variable KS2_MATSCORE will be used from the 2018/19 Key Stage 2 NPD. And similar to the primary outcome measure, the EVERFSM _6_P variable from the spring term school census 2016/17 will be used from the NPD.

The above primary and secondary analyses will then be repeated for the entire cohort rather than just EVERFSM pupils, thus forming another two secondary outcomes.

Two further attainment outcomes of reading and maths scaled scores will be used as secondary outcome measures. These will be for the entire year 3 cohort in 2016/17, who will take Key Stage 2 national curriculum tests in 2019/20. Variables KS2_READSCORE and KS2_MATSCORE will be used as the outcome measures from the 2019/20 Key Stage 2 NPD. This longer-term follow-up of pupils that were in year 3 in 2016/17 will be used to measure the impact of the intervention over three years: the intervention includes three years of support for schools to reach Gold Award level.

Non-attainment secondary outcomes will be collected from a social skills questionnaire. The questionnaire was administered by schools at baseline in the spring term 2016/17 and will be administered by NFER test administrators in the summer term 2018/19. This will include pupils who were in year 4 at baseline and in year 6 at end-point (the same cohort as the primary outcome measure). This questionnaire is an adapted version of a questionnaire used by Durham University as part of a separate evaluation of P4C (funded by the Nuffield Foundation). Each item included in this questionnaire represents the best single item available from a range of established instruments. Therefore, although it is usually good practice to use a collection of items that have already been demonstrated to be a reliable measure of the trait in question, it is not advisable to combine these items. Instead, they are recommended for use as stand-alone items by the instrument developers.

The chosen items should be the ones that are most closely matched to the theory of change model including ‘the 4Cs’: caring, collaborative, creative and critical thinking, and improvements in self-esteem, resilience, confidence and behaviour including tolerance and relationships. Previous research (Siddiqui et al., 2017) highlighted that there are two items that best reflect the aims of P4C. These are items 1A (‘I am good at explaining my ideas to other people’) and 1C (‘I can work with someone who has different opinions’) from the current questionnaire. There are five response categories for these items (ranging from 1 (not at all true) to 5 (completely true)) and they will be used as continuous variables in the analyses. These items will be labelled as ‘social and communication skills’ (item 1A) and ‘team work and resilience’ (item 1C) in our reporting and will be used as non-attainment secondary outcome measures for this trial.

## Analysis

The trial analysis will follow EEF’s Analysis Policy[2].

[2] https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Research_Report/2015_Analysis_for_EEF_evaluations.pdf

### *Primary intention-to-treat (ITT) analysis*

The primary outcome analysis will be 'intention-to-treat' and will only include pupils that are eligible for free school meals (as measured by EVERFSM_6_P variable). Multilevel models with two levels (school and pupil) will be used for the analysis to account for the cluster randomisation. We will use the R package for analysing education trials (eefAnalytics) to conduct our analysis.

The primary outcome measure KS2 reading scaled score (Variable KS2_READSCORE) will be the dependent variable with the following covariates:

- An indicator of whether the pupil was in an intervention school at baseline (reference category = in a control school)
- KS1 score in reading (as measured by KS1_READPOINTS variable from the 2014/15 NPD) as a prior attainment measure
- A dummy variable to identify the randomisation block (0 = January randomisation, 1 = March randomisation)

For the purposes of cross-study comparison, we will also report the point estimate (but not the confidence interval) from a model that does not include the dummy variable identifying the randomisation block.

### *Imbalance at baseline for analysed groups*

As the primary outcome is available from the National Pupil Database, it is anticipated that the level of missing will not exceed 5% at either the school or pupil level. No analysis of imbalance is therefore planned aside from the baseline comparison table specified in the report template.

### *Missing data*

As the primary analysis uses administrative data, it is anticipated that the number of pupils missing will be very small and so these cases can be excluded from the analysis without risk of bias. It is anticipated that the level of missing will not exceed 5% at either the school or pupil level so no missing data analysis is planned. If it does exceed 5%, reasons for missingness will be explored.

In the event that more than 5% of cases are found to be missing for a possibly biased reason, some further analysis will be carried out. In particular, a logistic multilevel model of whether or not an individual is missing, regressed on the prior attainment measure, dummy variable to account for the randomisation being conducted in two time-point blocks, the group allocation and further background NPD variables that are available from the standard school census supply (such as gender and age in months). This will help determine the extent of bias.

Missing data generally presents a problem for analysis, whether a pupil is missing a value for an outcome variable or for covariates (e.g. prior attainment). If outcome data is 'missing at random' given a set of covariates then the analysis has reduced power to detect an effect. If data is 'missing not at random' (for example, differential dropout in the intervention and control groups for unobserved reasons) then omitting these pupils, as with the primary 'completers' analysis, could bias the results. Imputing missing data could improve the robustness of the analysis and examine how sensitive the results are to alternative assumptions. It can also signal missing not at random if the imputed result is much different from completers analysis. Likelihood based methods (e.g. nlme function in R) are usually consistent with the results from multiple imputation if the missingness mechanism is missing at random. If it is not, some sensitivity analysis, for example using extreme values, may be necessary.

### *Non-compliance with intervention*

Fidelity analysis will be carried out on the primary outcome and the 2020 reading analysis only. We will incorporate fidelity information from SAPERE's awarding scheme to categorise schools according to how far through the 'Going for Gold' programme schools have progressed at summer 2019 and summer 2020. The table below shows the categories, derived from the SAPERE bronze/silver/gold awarding scheme. Schools will be assessed against the criteria when they make an application for an award. To ensure every school has a measure of fidelity for the analysis, schools that have not recently submitted an application for an award will be assessed against the criteria by a SAPERE trainer. By summer 2019 SAPERE expect intervention schools to have done the bronze award and meet 50% of the criteria for the silver award, and by summer 2020 to have done the silver award and meet 50% of the criteria for the gold award. The data that we receive from SAPERE will be based on this awarding scheme and can be summarised as below. Each awarding category will be converted to an engagement level as indicated in Table 2.

**Table 2: Categories from the SAPERE bronze/silver/gold awarding scheme**

| **Summer 2019** | | **Summer 2020** | |
|---|---|---|---|
| **Category** | **Engagement level** | **Category** | **Engagement level** |
| Bronze + 50% Silver, or above | 3 | Silver + 50% Gold, or above | 4 |
| Bronze, but not at 50% Silver | 2 | Silver, but not at 50% Gold | 3 |
| Below Bronze but some P4C activity occurred | 1 | Bronze, but not Silver | 2 |
| School withdrew before P4C activity started | 0 | Below Bronze but some P4C activity occurred | 1 |
| | | School withdrew before P4C activity started | 0 |

In order to obtain a more accurate measure of the 'pure' dosage effect of the intervention on pupil attainment the CACE impact estimate will be calculated. Because schools may potentially have unobserved characteristics that have an influence on both compliance with the trial and academic attainment 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 engagement level 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 outcome on the covariates used in the main model and will also include a covariate representing the pupil's estimated engagement level from the first stage of the model and an interaction term between the estimated engagement and the pupil's pre-intervention treatment allocation. The coefficient of the interaction term is the CACE estimate of the engagement 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.

A further factor that must be taken into account is the hierarchical nature of the data. To ensure that this factor of the data is accounted for correctly the R package ivpack, which has the functionality to correctly handle hierarchical data when using instrumental variables, will be used to perform the CACE analysis.

### Secondary outcome analyses

Attainment measures

Secondary outcomes of attainment will include three further measures from the Key Stage 2 NPD. These three models will be intention-to-treat multilevel models, will include all pupils and have an interaction term between the intervention indicator and the EVERFSM_6_P variable. These models will have similar covariates as the primary outcome models: an indicator for whether the pupil was in an intervention school at baseline, dummy variable to account for the randomisation block, and appropriate prior attainment measure. The dependent variables with corresponding prior attainment measures are listed below:

- For the first model, Key Stage 2 maths scaled score (KS2_MATSCORE variable from the 2018/19 NPD) will be the dependent variable and KS1_MATPOINTS variable from the 2014/15 NPD,
- for the second model, Key Stage 2 reading score (KS2_READSCORE variable from the 2019/20 NPD) will be the dependent variable and KS1_READ_OUTCOME variable from the 2015/16 NPD, and
- for the third model, Key Stage 2 maths score (KS2_MATSCORE variable from the 2019/20 NPD) will be the dependent variable and KS1_MATH_OUTCOME variable from the 2015/16 NPD.

Non-attainment measures

Two non-attainment measures (‘social and communication skills’ and ‘team work and resilience’) will be analysed using pupil questionnaire data available at end-point. These measures will be the dependent variables in two separate multilevel models containing two levels (school and pupil). Similar to the primary outcome measure, these models will only include EVERFSM pupils. The covariates that will be entered into each model will be an indicator of whether the pupil was in an intervention school at baseline, the equivalent social skills measure at baseline (a continuous variable) and a dummy variable to account for the randomisation block.

It is anticipated that these models might be affected by some level of attrition experienced at end-point pupil questionnaire administration. Therefore, multilevel multiple imputation will be used if appropriate as a sensitivity check that any missing outcome data is indeed missing at random, where the baseline pupil questionnaire data is considered to be the definitive dataset for analysis.

### Subgroup analyses

Sub-group analyses on the primary outcome will be carried out as per the protocol and the most recent EEF analysis guidelines. All sub-group analysis models will have KS2 reading scaled score (variable KS2_READSCORE from the 2018/19 NPD) as the dependent variable. As per the protocol, we will follow the primary analysis with an all-pupil analysis.

This will be the same model as the primary outcome model with all pupils included in the model rather than EVERFSM pupils.

This all-pupil analysis will be followed by interaction models exploring differential intervention effect on pupils from specific sub-groups. In these models, variables of interest will be interacted with the intervention variable. Given that further subgroup analysis performed by Gorard et al. (2015) indicated improvements in cognitive ability test scores for pupils for whom English is an additional language (EAL), EAL-by-intervention interaction will be investigated using LanguageGroupMajor variable. Given a possible differential impact of P4C on children of different abilities, we will include an interaction between intervention and prior attainment using KS1_READPOINTS variable. In order to explore differential intervention impact based on pupil FSM, we will include an interaction term between intervention and EVERFSM using EVERFSM_6_P variable. The effect of these interactions will be explored in a single model, rather than separate models for each interaction.

The interaction models will be similar to the all-pupil analysis where the covariates will also include the variable of interest along with the interaction term. E.g. the EVERFSM_6_P interaction model will have KS2_READSCORE as the dependent variable with the following covariates:

- an indicator of whether the pupil was in an intervention school at baseline (reference category = in a control school)
- KS1 score in reading (as measured by KS1_READPOINTS variable from the 2014/15 NPD) as a prior attainment measure
- EVERFSM_6_P variable
- EVERFSM_6_P * intervention
- Indicator for randomisation block

Data manipulation will be carried out in SPSS while the multilevel models will be run using the R package eefAnalytics.

### *Effect size calculation*

The numerator for the 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 a multilevel model, 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. We will use the R package for analysing education trials (eefAnalytics) which employs a slight correction to the calculation of effect size for cluster trials as referenced in the analysis guidance.

## Report tables

All the tables will be structured according to the EEF trial report template[3].

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

## References

Angrist, J.D. and Imbens, G.W. (1995). ‘Two-stage least squares estimation of average causal effects in models with variable treatment intensity’. *Journal of the American Statistical Association.* 90**,** 430, 431-442. [online]. Available: http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1995.10476535 [19 May, 2017].

Gorard, S., Siddiqui, N. and See, BH. (2015) Philosophy for Children Evaluation report and Executive summary. Education Endowment Foundation: London. [Available online]

Siddiqui, N. and Gorard, S. and See, B.H. (2017) 'Non-cognitive impacts of philosophy for children.', Project Report. School of Education, Durham University, Durham. Available: http://dro.dur.ac.uk/20880/1/20880.pdf?DDD34+DDD29+czwc58+d700tmt [21 July, 2017].

Stuart, E.A., Cole, S.R., Bradshaw, C.P. and Leaf, P.J. (2010). ‘The use of propensity scores to assess the generalizability of results from randomized trials’, Journal of the Royal Statistical Society, 174, 2, 369–386.