# Evaluation of the Peer Assisted Learning Strategies for Reading UK (PALS-UK) intervention, a two-armed cluster randomised trial. Statistical Analysis Plan

Evaluator (institution): Manchester Metropolitan University
Principal investigator(s): Cathy Lewin and Stephen Morris
Template last updated: August 2019

| | |
|---|---|
| **PROJECT TITLE** | **Evaluation of the Peer Assisted Learning Strategies for Reading UK (PALS-UK) intervention, a two-armed cluster randomised trial** |
| **DEVELOPER (INSTITUTION)** | Dr Emma Vardy (Nottingham Trent University) and Dr Helen Breadmore (University of Birmingham), supported by Dr Luisa Tarczynski-Bowles (Nottingham Trent University) |
| **EVALUATOR (INSTITUTION)** | Manchester Metropolitan University |
| **PRINCIPAL INVESTIGATOR(S)** | Cathy Lewin and Stephen Morris |
| **SAP AUTHOR(S)** | Sandor Gellen and Stephen Morris |
| **TRIAL DESIGN** | Two-arm cluster randomised controlled trial with random allocation at school level |
| **TRIAL TYPE** | Efficacy |
| **PUPIL AGE RANGE AND KEY STAGE** | 9-10 years old, KS2 |
| **NUMBER OF SCHOOLS** | 114[1] |
| **NUMBER OF PUPILS** | 4840 |
| **PRIMARY OUTCOME MEASURE AND SOURCE** | Reading attainment (New PiRA Summer 5 Test) |
| **SECONDARY OUTCOME MEASURE AND SOURCE** | Reading comprehension (WIAT-III UK)<br>Oral reading fluency (WIAT-III UK, Multi-dimensional fluency scale<br>Reading self-efficacy (Feelings about reading questionnaire)<br>Motivation for reading (Feelings about reading questionnaire) |

## SAP version history

| VERSION | DATE | REASON FOR REVISION |
|---|---|---|
| 1.0 [*original*] | 08/12/2022 | N/A |

[1] Numbers show the sample size at randomisation and are higher than the estimates in the original protocol written at the design stage

# Table of contents

SAP version history ..... 1
Table of contents ..... 2
Introduction ..... 3
Design overview ..... 3
Selecting the WIAT-III UK-T and MDFS sub-sample ..... 6
Baseline data ..... 7
Randomisation ..... 8
Sample size calculations overview ..... 8
Analysis ..... 10
Primary outcome analysis ..... 10
Secondary outcome analysis ..... 11
Subgroup analyses ..... 13
Additional analyses ..... 13
Imbalance at baseline ..... 13
Missing data ..... 14
Compliance ..... 16
Intra-cluster correlations (ICCs) ..... 17
Effect size calculation ..... 17
References ..... 19

# Introduction

This statistical analysis plan describes the proposed analysis of data from a cluster randomised controlled trial (CRCT) designed to evaluate the effectiveness of the Peer Assisted Learning Strategies for Reading UK (PALS-UK).

PALS is a whole-class, structured paired reading intervention. The version of the programme evaluated here is PALS-UK, based on the PALS grades 2-6 programme (intervention initially developed in the United States), with materials and training adapted and modernised for the UK context. PALS-UK is an ideal candidate for evaluation as it exemplifies a number of key elements of peer tutoring interventions: the provision of training to teachers and pupils and the use of structured activities to support high quality peer interactions. Peer interventions are also known to be low cost (which could make scale-up more feasible) and have been found to generate moderate/high effect sizes (EEF, 2018; Topping et al., 2011). This efficacy trial aims to provide necessary evidence on the impact of PALS-UK.

The PALS-UK programme will be delivered to Year 5 pupils over a total of 20 weeks in the school year 2022/2023. For the first four weeks children are trained on the PALS-UK activities, then in the following 16 weeks engage in self-directed learning. Pupils work in pairs, taking turns as coach and reader as they engage with four activities: partner reading, re-tell, paragraph shrinking and prediction relay. Sessions last 35 minutes and are conducted three times a week. According to the logic model, repeated reading with peer feedback will support all aspects of fluency: accuracy, automaticity and prosody, while the tasks of re-tell, paragraph shrinking and prediction relay will support reading comprehension. Taken together, it is predicted that the intervention will develop pupils' fluency, self-efficacy in reading, motivation for reading, reading comprehension and reading attainment.

Further details of the intervention including its theory of change can be found in the published trial protocol (Ainsworth et al., 2022).

# Design overview

The impact evaluation is designed to answer the following research questions:

### Primary research question

1. What is the difference in the average score for reading attainment among Year 5 pupils in schools exposed to PALS-UK, compared to Year 5 pupils in control schools exposed to business as usual conditions?

### Secondary research questions

1. What is the difference in the average score for oral reading fluency (rate) among Year 5 pupils in schools exposed to PALS-UK, compared to Year 5 pupils in control schools exposed to business as usual conditions?
2. What is the difference in the average score for reading fluency (multi-dimensional) among Year 5 pupils in schools exposed to PALS-UK, compared to Year 5 pupils in control schools exposed to business as usual conditions?
3. What is the difference in the average score for reading comprehension among Year 5 pupils in schools exposed to PALS-UK, compared to Year 5 pupils in control schools exposed to business as usual conditions?

4. What is the difference in the average score for reading self-efficacy among Year 5 pupils in schools exposed to PALS-UK, compared to Year 5 pupils in control schools exposed to business as usual conditions?
5. What is the difference in the average score for motivation for reading among Year 5 pupils in schools exposed to PALS-UK, compared to Year 5 pupils in control schools exposed to business as usual conditions?

## EXPLORATORY RESEARCH QUESTIONS

1. What is the difference in the average score for reading attainment among pupils who are entitled to Free School Meals (FSM) in schools exposed to PALS-UK, compared to the FSM pupils in control schools exposed to business as usual conditions?
2. What is the difference in the average score for reading attainment among pupils with special educational needs (SEND) who are in schools exposed to PALS-UK, compared to the pupils with SEND in control schools exposed to business as usual conditions?[2]
3. What is the difference in the average score for reading attainment among pupils scoring in the lowest quartile on the baseline New PIRA test in schools exposed to PALS-UK, compared to the pupils scoring in the lowest quartile on the baseline New PIRA test in control schools exposed to business as usual conditions?
4. What is the difference in the average score for reading attainment among pupils for whom English is another language (EAL) and whose score falls in the lower half of the sample distribution on the baseline New PIRA test in schools exposed to PALS-UK, compared to the same subgroup of Year 5 pupils in control schools exposed to business as usual conditions?

This is a pragmatic two-arm parallel stratified CRCT with whole schools allocated at random to treatment and control conditions on a 1:1 basis. The intervention is delivered to participating state primary schools in three English school commissioner regions: The North, the East Midlands and Humberside, and the West Midlands. In order to achieve balance on key school level covariates, randomisation was stratified by schools size and proportion of the school roll that were eligible for free school meals (FSM). The study population comprises pupils in trial schools entering Year 5 at September 2022. The primary outcome is the unstandardised score obtained by pupils in the New PIRA Summer 5 Test [3] to be sat in the summer of 2023. Pupil assessments will be administered by the evaluation team face-to-face in schools. Secondary outcomes for pupils are

- The scores obtained from WIAT-III UK-T[4]: reading comprehension and oral reading fluency subtest
- The scores obtained from the Multi-dimensional Fluency Scale[5]

[2] Some pupils with SEND who - based on the school's judgement - were unable to complete the PiRA at baseline will be excluded from the trial prior to randomisation. Therefore, this research question only applies to those SEND pupils who stay in the trial.

[3] https://www.risingstars-uk.com/series/assessment/rising-stars-pira-tests

[4] https://www.pearsonclinical.co.uk/store/ukassessments/en/Store/Professional-Assessments/Academic-Learning/Comprehensive/WIAT-III-UK-for-Teachers/p/P100009239.html

[5] See Rasinksi (2004) Available at https://files.eric.ed.gov/fulltext/ED483166.pdf

- The scores obtained from the Feelings about Reading questionnaire (measures reading self-efficacy and motivation for reading)[6]

The primary outcome measure (New PIRA Summer 5 Test) and the secondary outcome measure of reading self-efficacy and motivation (Feeling about Reading questionnaire) will be collected from all pupils within range of the intervention (Year 5 pupils). The remaining secondary outcome measures (WIAT-III UK-T and MDFS) will be collected from a subset of randomly selected pupils in each school. The effects of the intervention on the primary outcome will be estimated for four subgroups: 1) ever-FSM (using the variable EVERFSM_6), 2) designated SEND; 3) pupils scoring in the lowest quartile on the baseline New PIRA test, and (4) pupils with English as an additional language (EAL) and low prior attainment (i.e., pupils scoring in the lowest half on the baseline New PIRA test).

| | | |
|---|---|---|
| **Trial design, including number of arms** | | **Two-arm, stratified and cluster-randomised trial at the school level** |
| **Unit of randomisation** | | Schools |
| **Stratification variables** (if applicable) | | School size (one-form per year group versus two or more forms per year group)<br>Proportion of year group that are currently free school meals (split across the median sample proportion) |
| **Primary outcome** | variable | Reading attainment |
| | measure (instrument, scale, source) | Reading attainment (New PIRA Summer 5 Test) |
| **Secondary outcome(s)** | variable(s) | Oral reading fluency (rate); Oral reading fluency (multi-dimensional); Reading comprehension; Reading self-efficacy; Motivation for reading |
| | measure(s) (instrument, scale, source) | WIAT-III UK-T: reading comprehension and oral reading fluency subtest<br>Multi-dimensional Fluency Scale<br>Feelings about Reading questionnaire (measures reading self-efficacy and motivation for reading) |
| **Baseline for primary outcome** | **variable** | Reading attainment |
| | measure (instrument, scale, source) | Reading attainment (New PIRA Summer 4 Test) |
| | **variable** | Reading attainment; Reading self-efficacy; motivation for reading |

[6] Feelings about Reading questionnaire; the first part, measuring reading-self-efficacy, is adapted from Carroll & Fox (2017). Available at https://doi.org/10.3389/fpsyg.2016.02056. The second part, measuring motivation for reading, adapted from the scale used in the previous trial, is pending publication (Vardy, Breadmore and Carroll, in prep).

| Baseline for secondary outcome | measure (instrument, scale, source) | Progress in reading attainment (New PiRA Summer 4 Test)[7]<br>Feelings about Reading questionnaire |
|---|---|---|

As will be explained below, sample estimates of average effects will be obtained from separate regression models for each primary and secondary outcome, where the outcome will be the dependent variable. Sample estimates of treatment effects on the primary outcome and the secondary outcomes of Reading fluency (rate), Comprehension and Reading fluency (multi-dimensional) will be adjusted through the inclusion of month of birth and prior attainment in reading at the New PIRA Summer 4 Test as a covariate in the regression model. For the outcomes of Reading self-efficacy and Motivation, estimates of treatment effects will be adjusted for month of birth and the baseline scores on the respective subscales of the Feelings about Reading questionnaire (measuring reading self-efficacy and motivation for reading).

### *Selecting the WIAT-III UK-T and MDFS sub-sample*

For valid administration of the WIAT-III UK-T and MDFS, Assessors require training. It is estimated that administration of the two WIAT-III UK-T subtests will take 30 minutes per pupil. These features of testing raise the costs of data collection. To limit costs and minimise administrative burden, a sub-sample of pupils were randomly selected prior to randomisation to undertake an individually administered reading comprehension and oral reading fluency subtest and the Multi-dimensional Fluency Scale at end of the study. First, a single class in each multi-form entry school was randomly selected: classes were allocated a random number from a uniform distribution to four decimal places, and within each multi-form entry school, a single class was selected on the basis of the random number they were assigned and on an ascending basis. This was followed by selection at random of 10 pupils in each selected class (once again, using another set of random number from a uniform distribution to four decimal places, and selecting pupils with the lowest number in each preselected class). In single form entry schools, 10 pupils were selected directly. To maximise response rates, two further selection criteria were applied:

- Only pupils with a valid baseline PiRA score have been included in the pre-selection process[8]
- Five further pupils have also been selected at random in each school within the same process using the random numbers already assigned (i.e., five pupils that followed in the ranking were selected). These pupils will be approached to complete testing if pre-selected pupils are not present in school on the day tests are administered. The 'reserved' pupils are ranked from '1' to '5' based on the unique random number generated during the selection process, and test administrators will replace missing pupils in a predefined order (i.e., one missing pupil to be replaced with 'Reserve 1', two missing pupils to be replaced with 'Reserve 1' and 'Reserve 2' etc).

At the time of selecting the sub-sample, all but two primary schools had provided the Year 5 class allocations – information that was required to select a single class within multi-form entry

[7] Scales for WIAT-III UK-T III and the Multi-dimensional Fluency Scale are not administered at baseline. Therefore, baseline scores for oral reading fluency (rate), reading fluency (multi-dimensional) and reading comprehension are derived from the baseline New PiRA assessment.

[8] Note because PiRA was administered at baseline prior to randomisation, completion of PiRA at baseline is not correlated with the outcome of school randomisation and thus with treatment. This means that selecting only from these pupils for completion of more intensive data collection at follow-up, depending on whether they completed a baseline PiRA, is unlikely to lead to bias.

schools. For the two outstanding schools, pre-selection of the sub-sample of students *prior to* the randomisation was not possible (at the time of drafting this SAP, the missing information has been received and sub-sample selection was completed for these remaining schools).

To minimise the possibility that schools may inadvertently focus resources and effort on the children who will complete the related subscales, the identity of the selected sub-sample will only be revealed on the day of administration.

Due to the sample selection processes applied for follow-up testing, selection probabilities vary across schools. For example, pupils in single form entry schools are more likely to be selected for testing than those in multi-form entry schools. Weights that correct for these different selection probabilities have been calculated and will be supplied with the archive data sets[9], though they are not applied in the analysis proposed here because they are not directly relevant for our intention to treat estimates.

### *Baseline data*

The New PiRA summer 4 test was administered at baseline together with the Feelings about Reading questionnaire. They were administered online in June and July 2022 to 114 schools, with a response rate of 93.16%. Administration was distributed across four weeks to spread the load on the online testing system through which New PiRA was accessed, given that it was a peak time for test administration. However, schools still reported challenges with administering the online version of the PiRA, with a resultant impact on completion rates and on individual pupil results in the most affected schools. The details given below are based on notes taken by the evaluation team during the testing window.

At least 23 schools reported technical problems administering the PiRA:

- Five schools reported that their access codes, which are needed to login and complete the test, were not working.
- Ten schools reported issues relating to connection with the online PiRA including not being able to enter text, the screen freezing and being logged out.
- There were ten individual cases of pupils' answers failing to save; in four of these cases Rising Stars was able to retrieve and save the answers, and in six cases the pupil was invited to resit any unsaved questions.
- There were reports of unstable connections including skipping questions, being slow to load, being unable to return to questions or move between questions.

Some of these issues were likely to have been related to the school's browser or Internet connection, and schools were offered advice based on Rising Stars guidance to try different browsers and clear the cache. Some schools were then able to access testing successfully, sometimes on a rearranged date, while others continued to experience problems. A small number of schools did mention shortcomings in their own hardware or Wi-Fi that they were aware of and that they felt created problems with PiRA administration. Five schools also had problems administering the online Feelings about Reading questionnaire as well as the PiRA.

Seven schools wrote to give us further feedback on their experiences of administering the PiRA. Three explicitly said that online testing was a new experience to which some pupils were able to adapt more quickly than others. There were also comments relating to difficulties

[9] Data will be archived at the end of the project by the Education Endowment Foundation's Archive Manager, FFT Education.

navigating (particularly scrolling), the character allowance in text boxes not allowing pupils to give enough detail in answers, and the time limit on the test leaving some pupils unable to complete.

Feedback indicates then that schools faced technical difficulties with the online PiRA, as well as challenges with online testing for primary school pupils more generally. As a result of these issues, three schools reported significant numbers of pupils underperforming and questioned the accuracy of their school's results. In at least three other schools, the technical problems led to significant numbers of answers being marked 'no response' or pupil responses missing altogether. For these reasons the evaluation team has decided on paper-based administration for follow-up testing in June and July 2023.

### *Randomisation*

Overall 5325 schools were approached by the Delivery Team. Out of those, 124 have signed the MOU but 10 schools subsequently withdrawn before randomisation. Randomisation took place on the 2nd September 2022 and included 114 schools which had baseline data and a Memorandum of Understanding signed by the headteacher or member of the school Senior Leadership Team (SLT). The Delivery Team were informed about the outcome of randomisation on the same day, and schools notified shortly after. Randomisation followed the process that was set out in the protocol. Schools were stratified into four blocks on the basis of proportion of FSM students (split across the median sample proportion) and school size (one-form per year group, two or more forms per year group) in order to achieve balance on these level covariates. In total 57 schools have been allocated to the treatment condition and 57 to control. Table 1 below outlines actual allocations by stratum.

| | | | Number of schools per stratum | | | |
|---|---|---|---|---|---|---|
| | N Schools | N Pupils | Low FSM/single-form | High FSM/single-form | Low FSM/multi-form | High FSM/multi-form |
| **Treatment arm** | 57 | 2450 | 16 | 11 | 12 | 18 |
| **Control arm** | 57 | 2390 | 16 | 10 | 13 | 18 |
| **Total** | **114** | **4840** | **32** | **21** | **25** | **36** |

## Sample size calculations overview

The Table below provides an assessment of statistical power. Minimum detectable effect sizes were calculated using the software PowerUp (Dong & Maynard, 2013).

| | Protocol stage | | Randomisation stage | | Randomisation stage (with 10 percent school-level attrition) | |
|---|---|---|---|---|---|---|
| | OVERALL | FSM | OVERALL | FSM | OVERALL | FSM |
| **Minimum Detectable Effect Size (MDES)** | **0.203** | **0.234** | **0.206** | **0.230** | **0.218** | **0.243** |
| level 1 (pupil) | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 | 0.7 |

| | | Protocol stage | | Randomisation stage | | Randomisation stage (with 10 percent school-level attrition) | |
|---|---|---|---|---|---|---|---|
| | | OVERALL | FSM | OVERALL | FSM | OVERALL | FSM |
| **Pre-test/ post-test correlations** | level 2 (class) | 0 | 0 | 0 | 0 | 0 | 0 |
| | level 3 (school) | 0 | 0 | 0 | 0 | 0 | 0 |
| **Intracluster correlations (ICCs)** | level 2 (class) | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |
| | level 3 (school) | 0.10 | 0.10 | 0.10 | 0.10 | 0.10 | 0.10 |
| **Alpha** | | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |
| **Power** | | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 |
| **One-sided or two-sided?** | | Two | Two | Two | Two | Two | Two |
| **Average cluster size for level 1 (per level 2 unit)** | | 20[10] | 5[11] | 19.36 | 5.22 | 19.36 | 5.22 |
| **Average cluster size for level 2 (per level 3 unit)** | | 1.30 | 1.30 | 1.42 | 1.50 | 1.42 | 1.50 |
| **Number of schools** | intervention | 60 | 60 | 57 | 57 | 51 | 51 |
| | control | 60 | 60 | 57 | 57 | 51 | 51 |
| | **total** | 120 | 120 | 114 | 114 | 102 | 102 |
| **Number of pupils** | intervention | 2280 | 570 | 2450 | 801 | 2205 | 721 |
| | control | 2280 | 570 | 2390 | 885 | 2151 | 796 |
| | **total** | 4560 | 1140 | 4840 | 1686 | 4356 | 1517 |

Minimum detectable effect sizes are computed based on the assumed sample sizes at protocol and again based on the actual sample as randomised. Assumptions for Type I and II error rates, the pre/post-test correlations, and the assumption of two-sided tests for statistical significance have been maintained in the two sets of calculations.

The assumptions made in the sample size calculations are justified as follows:

- The proposed protocol sample size is 120 schools. MDES calculations for the 'randomisation' scenario are based on the total number of schools agreed to participate, which was slightly below our recruitment target (114 schools). A third scenario is also presented that assumes a 10 per cent attrition rate at school level, which is expected (see previous studies, e.g., Jay et al., 2017).
- Type I and II error rates to be set at five and 20 per cent respectively is standard practice in EEF trials, as they represent acceptable long run rates of error associated with different hypotheses of interest.
- Randomisation to intervention and control on a 1:1 basis.
- Estimates of the correlation between PiRA raw score for Reading Attainment were obtained KS1-KS2 correlation in the NPD, as reported by Allen et al., 2018.
- A three-level clustered design (pupils nested in classes nested in schools) was used assuming intra-cluster correlation ICC 0.10 at the school-level and 0.05 per cent at the

[10] We report the harmonic mean here to account for varying cluster size. Calculations are based on the average class sizes in English primary schools in 2020/21 as reported by the National Statistics (see Table below).
[11] Based on the previous trial, we assume that 25% of the pupils will be eligible for free school meals.

class-level. The low class-level ICC of 0.05 is consistent with previous research that has a three-level design equivalent to that proposed here (Boylan et al., 2018; Jay et al., 2017). This is also in line with the widespread practice estimating class-level ICC as being half of what is found at the school-level within primary education (Demack, 2017). Typically, EEF studies with a two-level design assume an ICC at the school level of 0.20, which is conservative. We have reduced this due to clustering at the class level and because we have results from the previous PALS evaluation commissioned by EEF, which shows school level ICCs of 0.14 (Culora et al., 2022). Furthermore, a recently published EEF Research paper reports an ICC at the school-level of 0.10 at Key Stage 2 for both Maths and English subjects (Allen et al., 2018).

- The intervention delivery team had the capacity to deliver the intervention to around 120 schools. Assuming relatively low school level attrition, the trial is powered to detect an effect of around 0.2 on the population of interest

The average cluster sizes, which are the average number of classes by school and the average number of pupils by class, were calculated using harmonic means. In the protocol stage, this is based on data released by the National Statistics (containing information on school and pupil numbers in English primary schools in 2020/21). Similarly, the cluster size figures for the 'Randomisation' scenario are based on the harmonic means that were calculated using the as-randomised sample. Using harmonic mean is recommended by the authors of programs for sample size determination such as PowerUp, in order to take account of variable cluster sizes (Dong & Maynard, 2013). The following table reports average number of classes per school and average number of pupils per class as calculated for the as randomised sample. This shows that arithmetic and harmonic means do differ, which confirms the appropriateness of applying harmonic means:

| Type of mean calculation | Arithmetic mean | Harmonic mean |
|---|---|---|
| **Average no of classes** | 1.72 | 1.42 |
| **Average no of pupils per class** | 24.69 | 19.36 |

## Analysis

The analysis will proceed on the basis of the principle of intention to treat (ITT). That is, pupils are identified in the analysis as members of the intervention or control group on the basis of their school's allocation to intervention and control conditions at randomisation regardless of whether the school subsequently takes part in the intervention or not. Where schools leave the study after randomisation and ask that their data are deleted, records for the relevant pupils will be removed from the sample file. Approaches to assessing the consequences of sample loss and possible strategies for missing data are discussed below.

### *Primary outcome analysis*

The primary analysis seeks an estimate of the average effect of intention to treat (AITT), of the intervention, on scores obtained from the New PiRA Summer 5 Test for Year 5 pupils. The primary outcome is a measure of reading attainment derived from the new PiRA reading test. PiRA has high internal validity and test reliability (Cronbach's alpha between 0.75 and 0.92), face validity (it is written to follow the national curriculum guidelines) and concurrent validity; showing a strong relationship with national test scores, and has a high correlation with external measures of attainment (McCarty and Ruttle, 2018). The domains for this measure include

vocabulary, comprehension, summary, inference, prediction and structure. A sample estimate of AITT will be obtained from a multi-level linear regression model taking the following three-level form:

$$Y_{kji} = \beta_0 + \beta_1 T_k + \beta_2 X_{kji} + \beta_3 Z_{kji} + \beta_4 S_k + w_k + u_{kj} + e_{kji} \ldots\ldots [1]$$

Here , $Y_{kji}$ is the unstandardised score obtained by pupil '$i$' in class '$j$' and school '$k$' from their New PiRA Summer 5 Test. The variable $T_k$ will take the value one if the pupil is in a school randomised to the intervention, zero otherwise. The sample estimate of the parameter of $\beta_1$ is the estimate of AITT. $X_{kji}$ represents student $i$'s raw score in their New PIRA Summer 4 Test. $Z_{kji}$ is a measure of a pupil's month of birth obtained from the baseline demographic data for pupil '$i$' in class '$j$' and school '$k$', and $S_k$ is a collection of all school-level stratum variables. $w_k$ is a school-level random effect, $u_{kj}$ is a class-level random effect and $\varepsilon_{kji}$ a pupil level residual term.

The school level random effect is assumed to be distributed normally in the population with zero mean and variance $\theta^2$, the class level random effect similarly with variance $\tau^2$. If the variance of $e_{kji}$ is $\sigma^2$, then the two intraclass correlation coefficients at the school and class levels are:

$$ICC_k = \frac{\theta^2}{\theta^2+\tau^2+\sigma^2} \ldots\ldots [2]$$

$$ICC_j = \frac{\tau^2}{\theta^2+\tau^2+\sigma^2} \ldots\ldots [3]$$

Parameter estimates will be obtained using STATA v17 statistical software.

For the primary outcomes, three further analyses will be performed. The first form of sensitivity analysis involves a reduced regression model that takes the form of equation [1] above but with the pupil baseline measure of reading attainment excluded:

$$Y_{kji} = \beta_0 + \beta_1 T_k + \beta_2 Z_{kji} + \beta_3 S_k + w_k + u_{kj} + e_{kji} \ldots.. [4]$$

This specification permits us to assess the extent to which inclusion of the baseline PiRA test score as a covariate influences the precision of the sample estimates. The second form of sensitivity analysis will mirror the regression model used for the primary analysis set out at equation [1] but with the age standardised PiRA score obtained at follow-up as the dependent variable instead of the raw score. This specification will omit the month of birth covariate previously included but will otherwise remain as equation [1]. This second specification will enable us to assess how far age-standardisation may influence results. Third, floor and ceiling effects will be assessed using histograms.

Uncertainty for the treatment effects in each specification will be reported in the form of continuous p-values and frequentist 95% confidence intervals. Regression estimates for treatment effects will be converted to effect sizes consistent with Hedges' g, as discussed in the effect size calculation section.

### *Secondary outcome analysis*

The following secondary analyses are proposed: we will estimate the effects on the reading self-efficacy and motivation for reading outcomes for the full sample. Secondary analyses will also involve estimating the effects on the MDFS outcome and the WIAT-III UK-T outcomes for the subset of 10 pupils per school selected at random for more extensive testing. Sample

estimates of average causal effect for reading self-efficacy, motivation for reading, oral fluency and comprehension will be obtained from fitting regression models to the relevant data consistent with the Model [1] above, and using the same statistical procedures, where the dependent variables '$Y'$ are derived from the relevant scales at follow-up. As WIAT-III UK-T III and the MDFS are not administered at baseline, the covariate '$X'$ is derived instead from the baseline New PiRA assessment.

Models are outlined in the Table below.

| Dependent variable | Model | Outcome Measure | Further covariates | Sample |
|---|---|---|---|---|
| **Oral reading fluency (rate)** | Hierarchical linear model | WIAT-III UK-T | 1) Age in months<br>2) Reading attainment (from the New PIRA Summer 4 Test) | 10 pupils /school |
| **Oral reading fluency (multi-dimensional)** | Hierarchical linear model | MDFS | 1) Age in months<br>2) Reading attainment (from the New PIRA Summer 4 Test) | 10 pupils /school |
| **Reading comprehension** | Hierarchical linear model | WIAT-III UK-T | 1) Age in months<br>2) Reading attainment (from the New PIRA Summer 4 Test) | 10 pupils /school |
| **Reading self-efficacy** | Hierarchical linear model | Feelings about reading | 1) Age in months<br>2) Reading attainment (from the Feelings about reading test) | All pupils |
| **Motivation for reading** | Hierarchical linear model | Feelings about reading | 1) Age in months<br>2) Reading attainment (from the Feelings about reading test) | All pupils |

In total WIAT-III UK-T has five subtests or scales and we will use two of them: reading comprehension and oral reading fluency. The reading comprehension subtest provides a score based on responses to a range of literal and inferential comprehension questions. Oral reading fluency (rate) is measured through the average number of words read correctly per minute from the two passages read by the pupil (where total word count for the two passages minus the errors made is divided by the time taken to read the passages, and then multiplied by 60 to convert the measure into seconds). While psychometric testing of the WIAT-III-UK has yet to be reported, the reliability of the U.S. version of the WIAT-III test has been assessed using the split-half reliability method, with mean reliability coefficients range between 0.91 and 0.98. With regard to validity, correlation coefficients range between 0.60 and 0.82 (Burns, 2010).

While the WIAT-III UK-T oral reading fluency subtest provides a basic measure of fluency (number of words correct per minute), the additional measure, the Multidimensional Fluency Scale (MDFS) (Rasinski, 2004) produces scores ranging from 4 to 16 and provides a qualitative measure of fluency based on judgements of: expression and volume, phrasing, smoothness and pace. The MDFS has been shown to be a reliable and valid measure of fluency (Paige et al., 2014).

The motivation to read scale developed by Vardy et al. (in prep) has been shown to have high reliability (Cronbach's α = .83). The reading self-efficacy scale is adapted from Carroll and Fox's (2017) original version of the scale with minor revisions to the phrasing of a few items

and an additional item added to more directly link to the PALS-UK intervention. This has a Cronbach's α value of .90 (Vardy et al., in prep).

### *Subgroup analyses*

Subgroup analysis will examine the effect of AITT on reading attainment scores for the following pupils:

- ever-FSM (using the variable EVERFSM_6)
- designated SEND
- pupils scoring in the lowest quartile on the baseline New PiRA

It is important to note that – prior to randomisation - a small proportion of SEND pupils within the sample were withdrawn because their teacher judged administration of the PiRA to be inappropriate for these pupils (school n = 57). Therefore, the SEND subgroup analysis will not include all SEND pupils that were recruited for the study. In addition, further exploratory analysis will examine the effects of PALS-UK for EAL pupils. To do this, it is proposed that an indicator is created that combines the NPD-type binary measure of EAL with a pupil's raw score on the baseline reading assessment, where that score falls in the lower half of the sample distribution. In other words, a binary indicator is created at the pupil level and takes the value '1' if a pupil uses EAL and that same pupil's score on the baseline reading test falls below the median score for the sample, '0' otherwise.

For all subgroups we will conduct separate analyses to explore differential effects for each subgroup by including an interaction term in Model [1] above, comprising the relevant subgroup indicator interacted with the treatment dummy indicator. These will subsequently be converted to Hedge's g, as per EEF reporting standards. These models are built up from equation 1, and thus have the following form:

$$Y_{kji} = \beta_0 + \beta_1 T_k + \beta_2 X_{kji} + \beta_3 Z_{kji} + \beta_4 S_k + \beta_5 subgroup_{kji} + \beta_6 T_k * subgroup_{kji} + w_k + u_{kj} + e_{kji}$$

Estimates of the causal impact of PALS UK of reading attainment of the specified subgroups of pupils will be expressed as an effect size, consistent with Hedges g using the same equation reported below (see section on Effect size calculation), but where the numerator is $\beta_6$ divided by the pooled standard deviation calculated for the subgroup only.

For all primary, secondary and subgroup analyses, unadjusted means, adjusted means and confidence intervals, and ESs will be reported.

### *Additional analyses*

No additional analysis is planned. All primary, secondary and exploratory analyses are described above.

### *Imbalance at baseline*

We aim to compare the characteristics of intervention and control group schools and pupils as measured in the 'as randomised' and 'as analysed' samples separately. We will include all schools and pupils in the 'as randomised' sample that have not subsequently withdrawn from the study after randomisation. The 'as analysed' sample will be all pupils for whom we observe a PiRA New PIRA Summer 5 Test score at summer 2023.

Tabulations will be presented to compare counts and proportions (for categorical variables), as well as means and standard deviations (for continuous variables), for the 'as randomised' and 'as analysed' samples with the following variables:

- Gender
- Age in months
- FSM Pupils
- SEND Pupils
- EAL Pupils
- Pupils scoring in the lowest half on the baseline New PiRA test
- Pupils scoring in the lowest quartile on the baseline New PiRA test
- PiRA reading score at baseline
- PiRA reading score at follow-up

At the time of drafting this SAP, data on all individual pupil level characteristics as well as PiRA baseline scores have been received by the evaluation team. Descriptive analysis was completed after randomisation comparing pupils and schools allocated to the intervention conditions to those allocated to control, comparing pupil level characteristics as well as PiRA baseline scores. These analyses are reported in the Table below. As can be seen randomisation has resulted in two groups that are well balanced. We received data on most pupil characteristics. There are no missing values on gender, FSM, SEND or EAL status and we observed only one pupil missing information on age. Baseline PiRA returns were high with the overall response rate of 93.16% for the total sample, with no significant differences between intervention and control groups in terms of missingness (6.5 per cent in the intervention as compared to 7.2 per cent in the control arm). This provisional analysis suggests that challenge due to missingness is more likely to come from missing data at follow-up.

| | | Intervention | Control | Difference |
|---|---|---|---|---|
| **Schools** | | 57 | 57 | |
| **Classes** | | 99 | 97 | 2 |
| **Pupils** | | 2450 | 2390 | 60 |
| **Gender** | | | | |
| | **Male** | 1203 | 1171 | 32 |
| | **Female** | 1247 | 1219 | 28 |
| **Age in months** | | | | |
| | **Mean** | 116.52 | 116.48 | 0.04 |
| | **Standard deviation** | 3.57 | 3.63 | -0.06 |
| | **Missing** | 1 | 0 | 1 |
| **FSM Pupils** | | 801 | 885 | -84 |
| **SEND Pupils** | | 419 | 411 | 8 |
| **EAL Pupils** | | 477 | 579 | -102 |
| **PiRA reading score** | | | | |
| | **Observed** | 2291 | 2218 | 73 |
| | **Missing** | 159 | 172 | -13 |
| | **% Missing** | 6.5% | 7.2% | -0.7% |
| | **Mean** | 16.15 | 16.17 | -0.02 |
| | **Standard deviation** | 8.24 | 8.18 | 0.06 |

### *Missing data*

For the primary analysis, sensitivity tests will be carried out to assess whether missing data at follow-up leads to biased or imprecise estimates of $\beta_1$.

Missingness that occurred before randomisation is unlikely to cause bias in estimated treatment effects but can result in diminished sample sizes. As shown above, 7.2 per cent of PiRA baseline scores are missing, and the rate of missingness is relatively balanced across trial arms. Rates of missing data in all other covariates (i.e., age) and crucial variables used for the subgroup analysis are trivial. We will examine the extent to which missingness at baseline lead to a loss of power. However, at this point missingness in the as-randomised sample does not seem to be substantial.

Incomplete outcome data – however – may lead to a loss of power as well as biased estimates of $\beta_1$. For the primary analysis potential sources of missingness subsequent to randomisation are likely to include:

- Parents requesting that their children be removed from the study, and their data deleted, subsequent to randomisation
- Pupils leaving the school prior to the completion of the New PiRA Summer 5 Test
- Schools withdrawing from the evaluation and requesting all data supplied by them to be deleted
- Pupils not present on the day of the New PiRA Summer 5 Test and unable to supply outcome data

In the first screening stage, we will examine the type of missingness: i.e., whether data is missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR). This includes calculating and comparing the rate of missing data in the trial arms. If we find the level of missingness to be problematic – i.e., missingness exceeds five per cent in both control and treatment groups – we will assess if available baseline covariates explain missingness. A multilevel logistic regression model is proposed where a binary response variable captures whether the follow-up observation for a sample member is observed or otherwise. This is the so called drop out model. The following explanatory variables that might be associated with missingness will be included in the model as covariates (all measured at baseline): gender, FSM status, SEND status, EAL status, PiRA baseline score and school size. Covariates found to be significantly associated with missing PiRA scores (with a 95 per cent confidence interval) will be considered explanators of the presence (or absence) of the follow-up observation on the primary outcome.

If missing data on the PiRA test at follow-up appear to exceed 5 per cent in anyone arm of the trial and evidence from the drop out model appears to indicate missingness associated with included covariates, further sensitivity tests will examine the consequences of missing data in the primary outcome, for the sample estimates of AITT in the primary analysis using multiple imputation.

We will use multiple imputation using chained equation (mice) to impute missing values for each variable so affected in our analysis using a fully conditional specification. This is the main advantage of mice over other procedures. Multiple imputation will involve the following steps:

- Select cases from our sample file that have a baseline record on the PiRA test
- For each variable we wish to include in our analysis that suffers from missing data, we specify an imputation model – in this application the imputation model will contain all other variables than the target variable for which imputations are required in the imputation model
- We specify the number of data sets to be created through the iteration process initiated in the `mice` program

- We specify a random number seed, which will we set equal to the date on which the multiple imputation is conduct in numeric form DDMMYYYY.
- Following the creation of the imputed datasets we will run mixed effects linear regression models of the form indicated by equation [1] above on the imputed data sets and combine the results using Rubin's rule.

To perform multiple imputation and the following analyses we will use either the `mi impute chained` and `mi estimate` commands in STATA v17 or the R programmes `mice` (v.3.14.0) and `miceadds` (v.3.15-21).

The results of the combined analysis can be compared to those obtained in the primary analysis to determine the sensitivity of our results to missing data under the assumption of MAR. We propose that 10 imputed data sets are created. We will examine the suitability of imputation procedures by examining diagnostic measures and plots with adjustments made as necessary.

### *Compliance*

PALS UK is conceived of as a whole-school intervention, meaning that compliance is defined at the school level. If a school meets all compliance criteria then all pupils within the school are deemed compliers. For the purposes of CACE (compliance average causal effect) analysis we define a complying school as one where

- at least one teacher from an intervention school attends the initial training event[12]
- there is evidence from that teacher's school that one or more pupils completed all four weeks of training.

Attendance at the training events is recorded by the delivery team. The pupil training compliance data can come from either class teacher weekly logs, RA observations, or the survey completed during top-up training confirming completion of training by that point. Schools are considered compliant if they fulfil both compliance criteria[13]. Conversely, schools that only fulfil one of two criteria or none of the criteria will not be deemed compliant. CACE analysis will be performed using Instrumental Variables (IVs) on the basis of Two Stage Least Square (2SLS). The purpose of this analysis is to estimate the impact of PALS UK for pupils that comply by virtue of attending a compliant school.

If significant non-compliance is encountered (around 10 percent or more[14]), a binary compliance variable will be constructed (1=complied, 0=not complied). Whilst it is likely that some schools assigned to the intervention group will be non-compliant, it is not possible for schools allocated to the control group to be non-compliant and to participate in PALS UK. This means that we face a situation of possible one-sided non-compliance. CACE therefore can be interpreted as the average effect of treatment on the treated. The proposed analysis involves the estimation of two equations. First we estimate a compliance equation in which the binary compliance variable is the dependent variable with the treatment group indicator as an independent variable:

[12] Note it is not possible for teachers to attend the top-up training event unless they have first attended the initial training. It is also not possible for pupils to receive training in PALS unless their teacher or at least one teacher in the school has attended initial training.

[13] Compliance criteria and definitions were determined following extensive discussions with both the developers and EEF and are on balance felt most appropriate

[14] This cut-off was arrived based on judgement following discussions with the delivery team about the level of compliance required for non-compliance to have substantive consequences.

$$D_{kji} = \gamma_0 + \gamma_1 T_k + \varepsilon_{kji}$$

Here $D_{kji}$ is coded to '1' if a pupil attends a compliant school. From this model we can obtain a predicted probability of compliance for all pupils in the sample $\widehat{D}_{kji}$. This predicted probability is included in an impact regression of the following form, where the covariates are defined as previously:

$$Y_{kji} = \beta_0 + \beta_1 \widehat{D}_{kji} + \beta_2 X_{kji} + \beta_3 Z_{kji} + \beta_4 S_k + w_k + u_{kj} + e_{kji} \ldots\ldots [1]$$

The sample estimate of $\beta_1$ is interpreted as an estimate of the effect of the intervention on the outcome for those that comply, or an estimate of the average effect of treatment for those treated.

### *Intra-cluster correlations (ICCs)*

The ICCs used for the power calculations reported above account for intra-cluster correlations at both school-level and class level. This 3-level model (i.e., pupils nested in classes nested in schools) follows previous EEF trials (Boylan et al., 2018; Jay et al., 2017)[15] where the assumed ICCs are conservative estimates of between-school and between-class variances. Other trials (e.g. Gorard et al., 2017; Humphrey et al., 2020; O'Hare et al., 2019; Rudd et al., 2017) previously funded by EEF have typically ignored clustering at the class level. This seems in part because empirical evidence suggests that class-level ICCs are typically very low for this year group. Our decision not to ignore class level clustering is informed by EEF research which argues that failing to account for class-level clustering can negatively impact on trial sensitivity and statistical power (Demack, 2019).

The final report will report the school-level and class-level ICCs at protocol and at analysis stage based on the primary outcome measure. For the primary analysis, this includes a null model that will yield an estimate of the full unconditional ICC for the primary outcome.

### *Effect size calculation*

Estimates of the causal impact of PALS UK on the primary outcome of reading attainment will expressed as an effect size, consistent with Hedges g. The equation for Hedges g is written as:

$$g = \frac{\hat{\beta}}{\sigma} \times \left(\frac{N-3}{N-2.25}\right) \times \sqrt{\frac{N-2}{N}}$$

In our application $\hat{\beta}$ is the sample estimate of $\beta_1$ from the regression model in equation [1] above. $\sigma$ is the unconditional pooled standard deviation for the dependent variable in equation [1] calculated across different levels in the data. The two factors to the right in the equation above adjust for bias in small samples. Given the size of the sample available to us these factors will be trivial and will therefore be ignored. Note that we do not use population standard deviations to calculate the effect size. This is because our sample is not a random sample of schools selected from the population and inferences relate to the sample rather than the

[15] The Dialogic Teaching evaluation conducted by Jay et al. (2017) that used GL Progress Test in English, Maths and Science as outcomes and KS1 test score as the baseline covariate for Y5 pupils reports class-level ICCs between 0.01 and 0.04. Another EEF trial, the ScratchMaths evaluation (Boylan et al., 2018) involving all Y5 and Y6 pupils had similarly low ICC scores (0.01 and 0.02).

population of all schools. This is considered acceptable because the trial is an efficacy study, which attempts to test the intervention under ideal circumstances.

In order to obtain a 95 per cent confidence interval on the effect size, we will use bootstrap procedures over 5,000 cycles. This will enable us to construct a confidence interval based on an empirical distribution obtained from the observed data using re-sampling and free from parametric assumptions.

This approach will be used to calculate effect sizes for both primary and secondary outcomes.

## References


Ainsworth, S., Gellen, S., Lewin, C., & Morris, S. (2022) *Evaluation of the Peer Assisted Learning Strategies for Reading UK (PALS-UK) intervention, a two-armed cluster randomised trial. Evaluation Protocol.* Education Endowment Foundation.

Allen, R., Jerrim, J., Parameshwaran, M., & Thompson, D. (2018). *Properties of commercial tests in the EEF database*. London: EEF Research Paper, (001).

Boylan, M., Demack, S., Wolstenholme, C., Reidy, J., & Reaney, S. (2018). *ScratchMaths: evaluation report and executive summary*. London: Education Endowment Foundation.

Burns, T. G. (2010) 'Wechsler Individual Achievement Test-III: What is the "Gold Standard" for Measuring Academic Achievement?', *Applied Neuropsychology, 17* (3), pp. 234–236.

Carroll, J.M., & Fox, A.C. (2017). Reading Self-Efficacy Predicts Word Reading But Not Comprehension in Both Girls and Boys. *Frontiers* in *Psychology, 7*.

Culora, A., Dimova, S., Ilie, S, Sutherland, A., & Gilder, L. (2022). *Peer Assisted Learning Strategies – UK Evaluation Report.* London: Education Endowment Foundation.

Demack, S. (2017) *Statistical Analysis Plan for ScratchMaths.* London: Education Endowment Foundation.

Demack, S. (2019). *Does the classroom level matter in the design of educational trials? A theoretical & empirical review.* London: Education Endowment Foundation.

Dong, N., & Maynard, R. A. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness, 6*(1), 24–67. https://doi.org/10.1080/19345747.2012.673143

Education Endowment Foundation (2018). *The EEF Teaching and Evidence Toolkit.* London: Education Endowment Foundation. Retrieved from: https://educationendowmentfoundation.org.uk/evidence-summaries/teaching-learning-toolkit/peer-tutoring/

Gorard, S., Siddiqui, N., See, B.H., Smith, E., & White, P. (2017). *Children's University: Evaluation Report and Executive Summary.* London: Education Endowment Foundation

Humphrey, N., Hennessey, A., Ashworth, E., Frearson, K., Black, L., Petersen, K., ... & Pampaka, M. (2020). *Good Behaviour Game. Evaluation Report and Executive Summary.* London: Education Endowment Foundation.

Jay, T., Willis, B., Thomas, P., Taylor, R., Moore, N., Burnett, C., ... & Stevens, A. (2017). *Dialogic teaching: Evaluation report and executive summary.* London: Education Endowment Foundation.

McCarty, C. and Ruttle, K. (2018) *Progress in Reading Assessment Manual (Stage 2)*, Second Edition, Hodder Education

O'Hare, L., Stark, P., Cockerill, M., Lloyd, K., McConnellogue, S., Gildea, A., ... & Bower, C. (2019). *Reciprocal Reading: Evaluation Report.* London: Education Endowment Foundation

Paige, D.D., Rasinski, T., Magpuri-Lavell, T., & Smith, G.S. (2014). Interpreting the relationships among prosody, automaticity, accuracy, and silent reading comprehension in secondary students. *Journal of Literacy Research, 46*(2), 123-156.

Rasinski, T.V. (2004). Creating fluent readers. *Educational Leadership, 61*, 46–51

Rudd, P., Aguilera, A.B.V., Elliott, L., & Chambers, B. (2017). MathsFlip: *Flipped Learning. Evaluation Report and Executive Summary.* London: Education Endowment Foundation.

Topping, K., Millder, D., Thurston, A., McGavock, K., & Conlin, N. (2011). Peer tutoring in reading in Scotland: thinking big. *Literacy*, 45(1), 3–9.

Vardy, E.J., Breadmore, H.L., & Carroll, J.M. (under review). Measuring the will and the skill of reading: Validation of the self-efficacy and motivation to read scale.