Peer Assisted Learning Strategies UK
Statistical Analysis Plan

Evaluator: RAND Europe
Principal investigator: Sonia Ilie

Education Endowment Foundation

PROJECT TITLE[1]	Peer Assisted Learning Strategies UK (PALS-UK)
DEVELOPER (INSTITUTION)	Dr Emma Vardy, Nottingham Trent University; Dr Helen Breadmore, Coventry University; Prof Kristen McMaster, University of Minnesota; and Profs Doug & Lynn Fuchs Vanderbilt University
EVALUATOR (INSTITUTION)	RAND Europe
PRINCIPAL INVESTIGATOR(S)	Dr Alex Sutherland[2] (January 2019 – June 2019) Dr Ilie Sonia[3] (June 2019- July 2020) Dr Sashka Dimova (July 2020 – present)
SAP AUTHOR(S)	Dr Sashka Dimova; Dr Sonia Ilie
TRIAL DESIGN	Two -arm cluster randomised controlled trial with random allocation at school level
TRIAL TYPE	Efficacy
PUPIL AGE RANGE AND KEY STAGE	Whole-class intervention, focus of the evaluation on Year 5 pupils (9-10 years old); Key Stage 2
NUMBER OF SCHOOLS	89
NUMBER OF PUPILS	2,176
PRIMARY OUTCOME MEASURE AND SOURCE	Pupil reading attainment in Year 5 Source: PiRA
SECONDARY OUTCOME MEASURE AND SOURCE	Oral reading fluency; Reading comprehension; Source: WIAT-III; and Self-efficacy in reading Source: Reading Self-efficacy survey

SAP version history

VERSION	DATE	REASON FOR REVISION
1.1	March 2021	To capture changes to project design and timelines in response to the COVID-19 pandemic.
1.0 [*original*]		*N/A*

[1] Make sure that the project title here matches the title of the document and the protocol. Please ensure that there is an identification as a randomised trial in the title as per CONSORT requirements.
[2] Formerly RAND Europe, currently Behavioural Insights Team.
[3] Formerly RAND Europe, currently University of Cambridge.

Table of Contents

Introduction

The aim of this efficacy trial is to assess whether 'The Peer Assisted Learning Strategies UK' (PALS – UK) leads to improvements in the reading fluency and reading comprehension skills of children in Year 5. PALS-UK is the updated, UK edition of PALS Grade 2-6. It is a whole-class, structured paired reading intervention initially developed in the United States by researchers Prof Doug Fuchs and Prof Lynn Fuchs from Vanderbilt University[4].

Peer tutoring interventions are appealing due to their potential impact and low cost. PALS-UK is a strong candidate for evaluation as it provides structure for within-class, same-age peer interactions, which could make scale-up more feasible. As a result, this efficacy trial will provide necessary evidence on the impact of PALS-UK.

The PALS-UK programme has been delivered over a total of 20 weeks (4 weeks training, 16 weeks intervention) in Year 5 classes during the winter and spring terms in school year 2019/2020. Children have been working in pairs for 30-35 minutes, three times a week taking turns to act as a reader or coach in a set of structured activities. PALS-UK sessions comprise of four activities: (i) partner reading, in which the children take turns as reader and coach. The stronger ability reader reads aloud first, while the coach monitors their reading and prompts them to correct their reading errors using the PALS-UK check-it procedure. Then they swap roles and the second reader re-reads the same passage; (ii) re-tell, in which the second reader only re-tells the story read with prompts from the coach; (iii) paragraph shrinking, in which both pupils summarise paragraphs that they have read using up to ten words for five minutes each; and (iv) prediction relay, in which both pupils have to predict what will take place in the next half-page of a story and then continue reading another half a page to check whether their prediction came true or not. Pupils have been split into their reading pairs based on reading ability, as judged by the teacher. The class is split into higher ability and lower ability readers, and each pair consists of one higher-attaining and one lower-attaining reader. Pairs were swapped approximately every 4 weeks after the initial training phase, up to a total of 4 pair swaps in the 20 weeks of intervention (including training).

PALS-UK sessions were part of the class' main activities. It has been recommended that PALS-UK sessions replace guided reading or other supplementary literacy activities (such as whole class reading) that a school currently does, but this was subject to school preference.

Over the course of the 2019/2020 school year, PALS-UK delivery involved a number of activities for school staff. To enable effective delivery, school staff have had one day of face-to-face training delivered by a USA PALS trainer Prof. Kristen McMaster with the delivery team and a half day of top-up training delivered by the delivery team. During the initial training teachers also received a PALS-UK manual, which was updated by the delivery team to ensure it is suitable for children in English school. Teachers also received an USB stick with resources, activity packs for training children, cards and videos. In addition to the manual, schools have also been given a set of books, to ensure there is reading material suited to different ability levels. In addition to training and resources, intervention schools could access ongoing support at ad hoc basis. Support was offered to teachers in interventions schools through videos of how to deliver PALS-UK, emails with reminders to switch pairs and tips for book selection over the intervention period, and just-in-time support offered over the phone. Support was also offered through peer observations.

In the current trial, PALS-UK was delivered to 44 schools, with another 45 schools assigned to the control arm. In the current evaluation, the primary focus is the overall effect of the intervention package on the: (i) the reading attainment of pupils in Year 5 classes and on (ii) oral reading fluency and (iii) on

[4] Developed by Douglas Fuchs, Lynn Fuchs and others https://vkc.mc.vanderbilt.edu/frg/what-is-pals/. Updates for PALS-UK by Emma Vardy, Helen Breadmore and Kristen McMaster.

reading self-efficacy of the pupils who are in randomly assigned intervention groups compared to pupils in control settings.

The PALS-UK programme is led by Nottingham Trent University and Coventry University and and is independently evaluated by RAND Europe. The study is funded by the Education Endowment Foundation.

For more information regarding the study and its background and design, please refer to the study protocol.[5]

Design overview

Owing to changes to the design outlined in the updated study protocol the impact evaluation research questions were modified. The research hypotheses this trial is able to address are as follows:**Hypothesis 1 (Primary Outcome):** Year 5 pupils in randomly allocated primary schools participating in PALS-UK (intervention schools) will have higher levels of reading attainment than pupils in control schools by Autumn 2020, as measured by Progress in Reading Assessment (PiRA).

Hypothesis 2 (Secondary outcome): Year 5 pupils in randomly allocated primary schools participating in PALS-UK (intervention schools) will have higher self-efficacy in reading than pupils in the control schools by Autumn 2020, as measured by pupil self-efficacy in reading.

Hypothesis 3: Year 5 pupils in treatment schools for whom English is another language (EAL) will perform better on all outcomes that EAL pupils in control schools.

Hypothesis 4: Year 5 pupils in treatment schools who are entitled to Free School Meals will perform better on all outcomes than FSM pupils in control schools.

Hypothesis 5: Lower and higher ability Year 5 pupils in treatment schools will perform better on all outcomes than pupils of similar ability in control schools.

Trial design, including number of arms		Two-arm, stratified and cluster-randomised trial (schools are clusters)
Unit of randomisation		School
Stratification variables (if applicable)		Region (Midlands and North East) School size (single-form entry versus multiple-form entry)
Primary outcome	variable	Reading attainment
	measure (instrument, scale, source)	Progress in Reading Assessment (PiRA)[6]
Secondary outcome(s)	variable(s)	Oral reading fluency; and Self-efficacy in reading.
	measure(s) (instrument, scale, source)	1: WIAT-III: Oral reading fluency and reading comprehension subtest; and 2: Reading Self-efficacy survey[7]

[5] Dimova S., Sutherland A.,(2019). Trial Evaluation Protocol: Peer Assisted Learning Strategies for Reading UK, London: Education Endowment Foundation
[6] https://www.risingstars-uk.com/series/assessment/rising-stars-pira-tests
[7] 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*, 2056.

Baseline for primary outcome	**variable**	Reading attainment
	measure (instrument, scale, source)	PiRA
Baseline for secondary outcome	**variable**	1: Reading attainment[8] 2: Reading self-efficacy
	measure (instrument, scale, source)	1: PiRA 2: Reading self-efficacy survey

UPDATE: Given school closures in January 2021 and extended lockdown due to the COVID-19 outbreak, it was not possible to continue collecting secondary outcome data on oral reading fluency and reading comprehension using the WIAT-III. Data was collected in a small number of schools (n=22) on the WIAT-III reading comprehension subtest before school closures were implemented. Thus, the WIAT-III data collection listed above was not completed and the secondary outcome analysis based on the WIAT III was foregone. However, we feel it will be beneficial to report mean outcomes on the Reading comprehension WIAT- III subtest for the small number of schools who completed the subtest in December 2020. Further details on the rationale for why secondary outcome testing using the WIAT-III was cancelled is provided in the updated study protocol. The PALS-UK evaluation was designed and executed as a two-group parallel, stratified, cluster-randomised trial, with school as the unit of randomisation. In total 89 schools were recruited from the Midlands and North East of England. To ensure comparability of schools in the intervention arm and the control arm, we randomised school within regions and by school size (single-form vs multi-form entry), in order to ensure the study arms were balanced on geographical location and the overall size of the school.[9] For more information on how schools were allocated to the treatment condition, please see Randomisation.

Schools were assigned to either treatment (PALS-UK) or control (business as usual). All schools signing up had a 50:50 chance of being assigned to the treatment group within each geographical region. All teachers (ostensibly working) in Year 5 classrooms in treatment schools were eligible for and are receiving the intervention.

There is only one treatment condition in this trial: throughout the 2019/20 academic year schools will receive training and support to implement PALS-UK. For control schools, it will be business as usual during the school year 2019/20.

Intervention schools: If a school has been selected to deliver PALS-UK (i.e. if they are in the intervention condition), all Year 5 teachers at the school can have access to the PALS-UK manual, training, resources and support needed to deliver the intervention.

Multi-entry schools: In multi-entry schools all Year 5 classes take part in the intervention activities, however only one randomly selected class will participate in the evaluation data collection activities, which involve baseline and outcome testing. Classes required to complete testing were randomly selected before programme delivery and before school randomisation; this will be the Year 5 class in single-form entry schools. The selected Year 5 class who completed a reading attainment test at baseline will be asked to undertake outcome testing at the end of the project. This significantly reduces testing costs but has only a small effect on the power of the study.

[8] Normally a baseline measure similar to the outcome measure would be used; however, to minimise impact on schools and pupils, the secondary outcome analysis will use the same baseline measure as the primary outcome analysis without collecting additional data at baseline.

[9] Oakes, J. M. (2013). Effect identification in comparative effectiveness research. *eGEMs*, *1*(1).

Control schools: Multi-entry control schools will continue implementing business as usual reading practices and only one randomly selected Year 5 classroom will undertake baseline and post-trial outcome testing.

Selection of sub-sample of children

Around one third of children in the classroom will be randomly selected to undertake an individually administered oral fluency subtest at the end of the study. The sub-sample of Year 5 children will be selected by RAND within four weeks of receipt of baseline data from the Delivery team (which occurred on the 27th of November 2019). However, the sub-sample of selected children who will complete the individually administered comprehension and fluency test will be revealed only at the time of the post-trial outcome testing. This minimises the possibility that schools may inadvertently focus resources and effort on the children who will complete the individual fluency and comprehension test if that is known too far in advance.

Randomisation

Allocation to treatment and control schools was conducted on the 12th September 2019 and included 89 schools which had baseline data and a MoU signed by the headteacher/SLT. Randomisation took place after baseline testing was completed in all schools. Schools were notified of their allocation on the 13th September 2019.

As pre-specified in the protocol, randomisation was stratified by region and form entry type (single/multi-form entry). Strata were constructed from regions (geography-based strata, Midlands vs. North-East), and based on school size (defined by the number of classes within a school:(one form entry school vs. multi-form entry schools). The stratification was undertaken to ensure that schools from the same region as well as schools with higher numbers of pupils were allocated to treatment or control evenly. Having uneven numbers of schools with one and/or multiple number of classes would mean there is a higher probability that treatment or control groups would be unequal in terms of size. To deal with unequal treatment fractions we used the command randtreat and the option misfits(global) in Stata (version 15.1).[10]

Schools were randomly allocated to one of the two arms of the trial within each stratum. In total 45 schools were allocated to the control condition, while 44 schools were allocated to the intervention condition.

Table 1 below shows actual allocations by region and school size. In total 45 schools were allocated to the control condition, out of which 24 were one-form entry and 21 were multi form entry; 44 schools were allocated to the intervention (PALS-UK) condition with total of 22 one-form and 22 multi-form entry schools.

Table 1: PALS-UK randomisation results

		Control group			Intervention (PALS-UK) group		
		N One-form schools	N Multi-entry schools	N Total schools	N One-form schools	N Multi-entry schools	N Total schools
1	**Midlands**	**13**	**18**	**31**	**12**	**19**	**31**
2	**North East**	**11**	**3**	**14**	**10**	**3**	**13**
	Total	**24**	**21**	**45**	**22**	**22**	**44**

[10] Carril, A. (2017). Dealing with misfits in random treatment assignment. *Stata Journal*, 17(3), 652-667.

Measures

Primary outcome measure: reading skills

The primary outcome for this study is a standardised measure of reading skills of pupils based on the Progress in Reading Assessment (PiRA) test (Hodder Education), which is sold in the UK by Rising Stars (RS) Assessment (provider for primary schools). The test takes around 40 minutes to complete[11]. This test assesses *general reading skills* (specifically comprehension, inference, and language, structure and presentation).

The test for the primary measure is administered at two time points: at baseline and at endline (as the outcome measure). Baseline testing took place in the first two weeks of September 2019 (between 2nd and 12th September 2019), and was overseen by the delivery team. Post-trial outcome testing was initially scheduled to take place in the Summer term between May and June 2020. Due to the COVID-19 pandemic, planned outcome testing in the early Summer term 2020 was postponed. Outcome testing was instead carried out between November and December 2020, when the schools were re-opened.

Secondary outcome measure

Two secondary outcome measures were intended to be captured as part of this study in addition to the primary outcome measure:

1. Reading fluency and reading comprehension

Reading fluency and language comprehension was intended to be assessed using the oral fluency subtest of the Wechsler Individual Achievement Test–III (WIAT-III)[12,13]. The oral fluency subtest measures: speed, accuracy, prosody, and fluency in reading. Reading comprehension was intended to be assessed with the reading comprehension scale, which measures: comprehension of various types of text, including stories, advertisements and how-to passages[14]. For the oral fluency sub-test pupils are asked to read a passage aloud; in the reading comprehension test they are asked to read a passage silently. Then in both tests, pupils are asked to respond orally to questions.

To reduce the testing burden and costs, it was agreed to administer the WIAT-III subtests to ten randomly selected children per school (for more information on the selection process look at Selection of sub-sample of children). Pupils were randomly sampled at a later date. For multiple-entry schools, pupils were selected from within the randomly selected class who would complete both baseline and post-trial outcome testing. Within the constraints of the school day, it was intended that the WIAT-III will be completed on the same day with the PiRA test. In terms of order, the PiRA would have been completed first and there would have been break time before children completed the self-efficacy questionnaire and subsequently the WIAT-III. More information on the selection of the sub-sample is presented in *Selection of sub-sample of children.*

Update: Due to the COVID-19 lockdown and school closures in the spring and summer of 2020, testing based on the WIAT-III was planned to be administered over December 2020 and January 2021. To support a no visitor policy in schools in response to COVID 19, WIAT-III data collection was planned to be undertaken remotely. Testing using the WIAT-III started in December 2020 and one of the subtests (Reading comprehension) was completed in 22 schools, However, given school closures and the implementation of national lockdown measures in January 2021, the possibility to collect the WIAT-III was no longer feasible. As a result, it was not possible to continue collecting the secondary outcome

[11] https://www.risingstars-uk.com/series/assessment/rising-stars-pira-tests
[12] https://www.pearsonclinical.co.uk/Sitedownloads/Productpdfs/wiat-iii-uk-subtests-description.pdf
[13] Burns, T. G. (2010). Wechsler Individual Achievement Test-III: What is the 'Gold Standard' for Measuring Academic Achievement?. *Applied Neuropsychology*, *17*(3), 234-236.
[14] Prosody refers to reading expression; the ability to incorporate the rise and fall of pitch, to pause within and between sentences and etc. For more information look at: Schwanenflugel, P. J., Hamilton, A. M., Kuhn, M. R., Wisenbaker, J. M., & Stahl, S. A. (2004). Becoming a fluent reader: reading skill and prosodic features in the oral reading of young readers. *Journal of educational psychology*, *96*(1), 119.

measure on Oral reading fluency and Reading comprehension. Further details on the rationale for why secondary outcome testing using the WIAT-III was cancelled can be found in the updated study protocol.

2. Reading self-efficacy

The evaluation also aims to understand if PALS-UK positively affects pupils' reading self-efficacy. Self-efficacy refers to children's beliefs in their capability to produce satisfactory attainment[15]. A self-reported survey containing 20 items, developed by researchers from Coventry University and Nottingham Trent University, will be used to assess children's self-efficacy. The survey has been tested in a small study with thirty children and has an internal reliability of (α = 0.912). More information on the questionnaire is available in Carroll and Fox[16].

To assess changes in pupils' reading self-efficacy, the survey was administered at the same day as the primary outcome assessment. Please see above details about the time and order of administration.

Optional follow-up

Long-term follow-ups of educational interventions are rare; as such this trial could also offer a good opportunity to assess the effectiveness of PALS-UK measured one year after the end of the intervention, once pupils have transitioned into Year 6. This opportunity – likely using KS2 outcome data – is currently being considered by the EEF and thus not discussed further here.

Sample size and power calculations overview

The initial power calculations were based on both the information provided in the Invitation to Tender and the subsequent set-up meetings with the Delivery team and the EEF. Power and minimum detectable effect size (MDES) calculations were performed using PowerUp[17].

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM[18]
Minimum Detectable Effect Size (MDES)		0.228	0.263	0.231	0.269
Pre-test/ post-test correlations	level 1 (pupil)	0.7	0.7	0.7	0.7
	level 2 (class)	NA	NA	NA	NA
	level 3 (school)	0	0	0	0
Intracluster correlations (ICCs)	level 2 (class)	NA	NA	NA	NA
	level 3 (school)	0.13	0.13	0.13	0.13
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		Two	Two	Two	Two
Average cluster size		28	7	24.45	6.12

[15] Zimmerman, B. J. (2000). Self-efficacy: An essential motive to learn. *Contemporary educational psychology*, *25*(1), 82-91.
[16] 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*, 2056.
[17] Dong, N., & Maynard, R. (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.
[18] We do not currently have information on FSM status. This will be collected from the NPD prior to outcome testing; For this MDES we assume the same rate of FSM eligibility as at protocol stage.

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM[18]
Number of schools	intervention	45	45	44	44
	Control	45	45	45	45
	Total	90	90	89	89
Number of pupils	intervention	1260	315	1113	278
	Control	1260	315	1063	267
	Total	2520	630	2176	545

The power calculations assume equal allocation to treatment and control groups. The estimation at protocol stage is based on 90 schools. With one Year 5 class per school included in the evaluation, at protocol stage we assumed an average cluster size of 28 pupil (for more information on class selection in multi-entry form school see Multi-entry schools subsection under Design overview). Based on previous research, we assumed the proportion of variance in Level 1 outcomes explained by Level 1 covariates R_1^2 is 0.49 (equating to a correlation of 0.70, as per the power calculation table above) and R_2^2 of 0.00.[19] We used two-level clustered designs and base our calculations on an intra-cluster correlation (ICC) of 13 per cent. Assuming a desired power of 80 per cent, alpha of 5 per cent, an ICC of 0.13 and a continuous, normally distributed outcome, the protocol MDES was *d*=0.228. Using the parameters above and assuming that on average there are around seven FSM pupils in one Year 5 class (a rate of 25% FSM eligibility), the MDES was 0.263. As such, even though considered an efficacy trial, the study should be powered to detect meaningful differences between groups.

At randomisation, there were 89 schools and a total of 2,176 pupils with a completed baseline measure, the criterion for inclusion in the trial sample. This represents an average of 24.45 pupils per school. With the achieved sample and ICC of 13% the estimated MDES is *d*=0.231. As such, the MDES reported in the SAP is similar in size with the MDES in the study protocol (MDES of *d*=0.228).

Potential attrition issues

We expected that some attrition would be caused by COVID-19. Keeping all other parameters the same, we estimated the potential impact of different levels of attrition (at the school level and at the pupil level) on the minimum detectable effect size (MDES).

School-level (i.e. schools drop out, this also impacts pupil numbers)		Pupil-level (i.e. schools are all retained, but pupil numbers per school reduce)	
Attrition	MDES	Attrition	MDES
10%	0.244	10%	0.232
15%	0.250	15%	0.233
20%	0.259	20%	0.234
25%	0.267	25%	0.235
30%	0.278	30%	0.237
50%	0.329	50%	0.244

[19] Choudry, S., Squires, G., and Humphrey, N. (2017). *Statistical Analysis Plan for Achievement for All*. Retrieved from:https://educationendowmentfoundation.org.uk/public/files/Projects/Round_9_-_Achievement_for_All_SAP_(amended).pdf

Analysis

The outcome analysis will be carried out using an intention-to-treat (ITT) approach.[20] The analysis will include all randomised schools/pupils in the groups to which they were randomly assigned, regardless of the treatment actually received, withdrawal from the intervention post-randomisation, or deviations in programme implementation. This principle is key in ensuring an unbiased analysis of intervention effects. This approach compares outcome means for the treatment and comparison groups, and subjects are analysed according to their randomised group allocation. The ITT approach is inherently conservative as it captures the averaged effect of *offering* the intervention, regardless of whether the participants complied with assignment.

Primary outcome analysis

The primary outcome is pupil-level PiRA test scores, which are age-standardised to a mean score of 100, showing whether a pupil is above or below average as compared to PiRA's national standardisation sample. The age–standardised score can be used to compare pupils' reading ability against children of the same age.

To estimate the impact on the primary outcome we will use a two-level multilevel model to account for clustering of data. Multilevel approaches assume that the schools in the study are a random sample of all schools and that the multilevel modelling framework can flexibly handle complex variation within/between schools.[21,22,23]

The main analysis consists of the model for outcomes of pupils nested in schools, which is:

$$(1)\ Y_{ij} = \beta_0 + \mathrm{PALSUK}_j\tau + Z_j\beta_1 + X_{ij}\beta_2 + u_j + e_{ij}$$

where Y_{ij} is the PiRA score for child *i* in school j; β_0 is the cluster level coefficient for the slope of a predictor on language; PALSUK_j is a binary indicator of the school assignment to intervention [1] or control [0]; Z_j are school-level characteristics, here the two stratifying variables of geographical location and single/multi-entry form nature of the school (as used for randomisation); X_{ij} represents characteristics at pupil level (pupil $_i$ in school $_j$), specifically the baseline (pre-intervention) PiRA score; u_j are school-level residuals ($u_j \sim i.i.d\ N(0, \sigma_u^2)$) and e_{ij} are individual-level residuals ($e_{ij} \sim i.i.d\ N(0, \sigma_e^2)$).

Equation (1) is known as a 'random intercepts' model because $\beta_{0\mathrm{j}} = \beta_0 + u_j$ is interpreted as the school-specific intercept for school j and $\beta_{0j} \sim i.i.d\ N(\beta_0, \sigma_u^2)$ is random as in, it is a number that can take any value.

Our target parameter (i.e. the focal result of the trial) is τ, a binary treatment/control indicator variable. That will tell us the average effect of the intervention on pupil outcomes in treatment schools compared to those in control schools. This will be presented as a predicted, adjusted mean difference in outcome PiRA score between with two groups with a 95% confidence interval (CI) and p-value.

All analyses will be run in Stata, versions 15.1 onwards.

[20] Fisher, L. D., Dixon, D. O., Herson, J., Frankowski, R. K., Hearron, M. S.,et al. (1990). Intention to treat in clinical trials. In K. E. Peace (ed). *Statistical Issues in Drug Research and Development* New York: Marcel Dekker, pp. 331–350.
[21] Hox, J. (1998). Multilevel modeling: When and why. In Balderjahn I., Mathar R., Schader M. (eds) *Classification, Data Analysis, and Data Highways. Studies in Classification, Data Analysis, and Knowledge Organization.* Berlin, Heidelberg: Springer.
[22] Snijders, T. A. (2005). Power and sample size in multilevel modeling. *Encyclopedia of statistics in behavioral science*, *3*, 1570-1573.
[23] Snijders, T. A., & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological methods & research*, *22* (3), 342-363.

Secondary outcome analysis

The secondary outcome analyses will use the reading self-efficacy survey. Pupil's self-efficacy will be assessed following the same specification to equation (1) listed under primary outcome analysis above, but we will substitute the the self-efficacy score as secondary outcomes.

The vector of pupil-level characteristics in Equation (1), X_{ij}, will include the relevant baseline measure: the baseline reading self-efficacy score.

Update: Owing to changes outlined above, secondary outcome analysis on pupil's oral reading fluency and reading comprehension using the WIAT III subtests is not possible. However, we feel it will be beneficial to report mean outcomes for the treatment and control groups separately on the Reading comprehension WIAT-III subtest, for the small number of schools who completed the subtest in December 2020 (n=22). Given the size of the sample the analysis is not capable to detect relevant differences.

Subgroup analyses

As defined in the trial protocol we will also conduct the analysis for the following subgroups, using the same model as our primary analysis:

1. Children who are eligible for (FSM) as registered in the National Pupil Database (NPD) (using the variable EVERFSM_6); We will explore differential effects for FSM pupils as they are considered a key target group by the EEF. We will substitute an interaction term in the main equation (1) above, to account for the FSM subgroup and the treatment allocation while retaining the whole analytical sample in the model. As a robustness check we will also undertake the FSM sub-group analysis using the FSM-eligible sample only (i.e. as a separate sub-sample).

2. Children who are registered as EAL in the NPD. The EAL indicators identifies children who are routinely exposed to other languages in their home, and evidence[24] suggests it is not a good indicator of pupils' later attainment because of the heterogeneous group it captures, with English proficiency emerging as a better predictor of later attainment instead. However, the focus on EAL is important as an earlier evaluation of a PALS programme in 10 schools in the US, which focused on English Language Learners (ELL) pupils with Learning Difficulties (LD) found a large positive effect for this subgroup.[25] The same analytical approach as for the FSM sub-group analysis will be taken.

3. Children with lower reading ability who score below a threshold point on the PiRA age-standardised scores during the baseline test. Finally, we will explore if programme effects differ across high and low reading achievers as previous findings about the effects of pairing low with high ability pupils are mixed.[26] The evaluation of PALS-UK will inform if there are any differences in programme effect for lower reading ability pupils compared to higher ability pupils. In the analysis pupil ability will be based on the PiRA reading assessment scores at baseline, as outlined above. Low reading achievers will be identified as pupils scoring below average on the age standardised PiRA test. We will consult with the test developers on the definition of this threshold and construct a binary variable (below the threshold; at or above the threshold) to enter into an analysis of the same type as the FSM sub-group analysis. Additionally, we will undertake an additional exploratory analysis (see Additional analysis below) for the very low ability readers (bottom quartile of PiRA scores).

[24] Strand, S., & Hessel, A. (2018). English as an Additional Language, proficiency in English and pupils' educational achievement: An analysis of Local Authority data.
[25] Sáenz, L. M., Fuchs, L. S., & Fuchs, D. (2005). Peer-assisted learning strategies for English language learners with learning disabilities. *Exceptional children*, *71*(3), 231-247.
[26] Lou, Y. Abrami, P. C. Spence, J. C. Poulsen, C. Chambers, B. d' Apolonia, S. Within-class grouping: A meta-analysis Review of Educational Research 66 423–458 1996.

The study will report mean outcomes by sub-categories of, children eligible for FSM children with EAL and low reading achievers at baseline as a basic descriptive step.

We will estimate programme impact on the primary outcome for the subgroups above using separate models. Each respective subgroup indicator (FSM, EAL, low reading ability) will be incorporated into the regression analysis as a binary variable [1] if a child is identified as such and as [0] if they are not. Each of these indicators will then separately be interacted with treatment allocation to assess the conditional impact of PALS-UK on the respective sub-group pupils.

As these analyses are exploratory and likely underpowered, we will report point estimates and confidence intervals but will not report significance tests/p-values.

Additional analysis

We will carry out one single additional exploratory analysis: estimating the impact of PALS-UK on pupils with very low reading ability at baseline, defined as the bottom quartile on the baseline PiRA score.

The estimation approach will mirror the approach for the subgroup analysis, substituting the sub-grouping variable in the interaction term with treatment with a very low ability variable, taking the value 1 if the pupils scores below the 25th percentile on the baseline PiRA test and 0 otherwise.

Balance at baseline

We have taken an active approach to address any potential imbalance by stratifying at randomisation. A well-conducted randomisation, in expectation, yields groups that are equivalent at baseline[27]. Because schools are randomly allocated to the control and intervention conditions, any imbalance at baseline will have occurred by chance. To check for, and monitor, imbalance at baseline in the realised randomisation, baseline equivalence will be conducted at the school and pupil level.

At the school level, we will check the balance in the following variables by means of cross-tabulations and histograms that assess the distribution of each characteristic within the control and intervention groups:

- Proportion of children in Year 5 class speaking English as an additional language (EAL).
- Proportion of children in Year 5 class eligible for FSM.
- Proportion of all children in the school eligible for FSM.
- KS2 reading fine grained scores.

At the individual (pupil) level, balance will be assessed for the following characteristics:

- EAL status.
- FSM status for pupils.
- Gender.
- Average age in months.
- Reading ability (PiRA baseline test).

At the time of drafting this SAP, data on three of the individual pupil level characteristics is available. The table below documents the balance between pupils in schools allocated to the control and intervention conditions respectively. We conclude that randomisation has resulted in intervention and

[27] Glennerster, R. and Takavarasha, K. (2013) *Running randomized evaluations: a practical guide.* London: Princeton University Press.

control groups that are balanced and comparable in terms of reading comprehension attainment and pupil characteristics (date of birth, gender).

Mean standardised PiRA scores at baseline displayed a balanced distribution. The mean PiRA score was 93.84 for the intervention and 94.46 for the control group. Both groups are balanced also in terms of age in months with a difference of 0.5 percentage point.

Baseline Pupils Characteristics	Total Sample (n=2,176)	
	Intervention Schools (Pupil N=1,113)	Control Schools (Pupil N=1,063)
Gender		
Male	542 (52%)	500 (48%)
Female	507 (48%)	536 (52%)
Missing	64	29
Observations with data	1,049	1,034
Age in months		
Mean	113.44	113.94
Standard deviation	(3.67)	(3.61)
Missing	44	31
Observations with data	1,069	1,032
PiRA age-standardised score at baseline		
Mean	93.84	94.46
Standard deviation	(16.59)	(16.63)
Missing	0	0
Observations with data	1,113	1,063

As data becomes available[28] on the other characteristics of interest at baseline (additional to above table), we will assess baseline equivalence at the school level and the pupil characteristics as defined above.

As above, we will not carry out statistical significance tests to assess balance at baseline, as the premise of statistical testing at baseline does not hold in randomised controlled trials[29]. Instead, tables of the means along with distributions (for continuous variables) or counts with percentages (for categorical variables) will be presented, as above[30,31].

Missing data

We will explore the extent of missingness, and then also explore the pattern of any identified missingness. The procedure outlined below refers to the primary outcome measure (and associated

[28] FSM and EAL status will be obtained from the National Pupil Database.
[29] Baseline Data. *Consort (2010)*. Retrieved 26 March (2019): http://www.consort-statement.org/checklists/view/32-consort/510-baseline-data
[30] There is a convention in some disciplines that a 10pp (or larger) difference in treatment and control means at baseline constitutes 'imbalance' is thus justification for including those measures in sensitivity analyses, but there are counter-arguments to this idea. See Roberts, C. and Torgerson, D. (1999) 'Baseline imbalance in randomised controlled trials', *BMJ*, 319:185; de Boer et al. (2015) 'Testing for baseline differences in randomized controlled trials: an unhealthy research behavior that is hard to eradicate', *International Journal of Behavioral Nutrition and Physical Activity*, 12:4.
[31] Senn, S. (1994) 'Testing for baseline balance in clinical trials', *Statistics in Medicine*, 13: 1715-1726.

primary analysis); for the secondary outcome measures we will report the extent of missing data but not undertaken any additional analyses.

Missing data can arise from item non-response or attrition of participants at school and pupil levels. Even though it is important to include all data, it can be problematic to apply the intention to treat principle if we are not able to complete follow up testing for all randomised schools. We propose the following steps for dealing with missing data.

Firstly, the ideal solution to the problem is to avoid missing data altogether. We will attempt to follow up with all randomised schools even if they withdraw from allocated treatment. Secondly, each test outcome measure mentioned above (PiRA, WIAT-III oral fluency and reading comprehension) comes with associated marking procedures that result in item non-response not being an issue (in the sense that a missing answer is considered an 'incorrect' answer, and therefore does not contribute to the final score). For the reading self-efficacy measure we will address any item non-response by using a mean rather than a sum score (in the absence of validation data to provide a different factor score derivation procedure).

If the outcome data are incomplete, we will first determine the proportion of missing data in the trial. We will explore attrition across trial arms as a basic step to assess bias.[32] We will provide cross-tabulations of the proportions of missing values on all outcome measures.

To assess whether there are systematic differences between those who do not provide a valid PiRA post-test score and those who do not – and thus whether these factors should be included in analysis – we will model missingness through a logistic regression model at follow-up as a function of baseline covariates, including treatment. The analysis model for this approach will mirror the multilevel model given above (pupils clustered in classes), but the outcome will be a binary variable identifying missingness (yes/no). This will also make use of the fact that baseline data relevant to the primary outcome analysis is complete for all trial participants. If any of the baseline covariates are seen to be statistically significantly associated with missing primary outcome data (at the 5% level) then the primary analysis will be repeated adjusting additionally for these covariates as fixed effects.

If there is less than 5% missingness overall (i.e., the primary analysis model includes at least 95% of randomised pupils), we will carry out a complete-case analysis, and undertake an exploratory robustness analysis using a full-information maximum likelihood (FIML) approach (instead of multiple imputation (MI)), because FIML can be estimated in a single model and simulation studies[33] show that it can reduce bias as well as MI[34].

If there is more than 5% missingness overall, we will undertake analysis to understand if the data appears to be missing completely at random (MCAR), or whether the weaker Missing at Random (MAR) assumption applies. We will assess if data is consistent with being missing completely at random (MCAR) by using the `mcartest` Stata package; this implements Little's test of MCAR. We would use this as an indication only, as opposed to a definitive conclusion, and supplement this by creating dummy variables to identify missingness for the primary outcome variable[35] and then explore (through simple t-tests) if this missingness indicator is associated with any of the other variables in the dataset, adjusting for multiple comparisons. This will allow us to understand if the missing data pattern is MCAR or, if the main outcome is missing conditional on other variables (MAR).

[32] Higgins, J., Altman, D., Gøtzsche, P., Jüni, P., Moher, D., Oxman, A., Savovic, J., Schulz, K., Weeks, l. and Sterne, J. Cochrane bias Methods Group, Cochrane Statistical MethodsGroup (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ,* 343**,** d5928.

[33] Allison, P. D. (2003). Missing data techniques for structural equation modeling. Journal of abnormal psychology, 112(4), 545.

[34] Multiple imputation is not necessarily the gold standard in missing data handling in RCTs, with other (simpler) methods providing similarly unbiased estimates: Sullivan, T. R., White, I. R., Salter, A. B., Ryan, P., & Lee, K. J. (2018). Should multiple imputation be the method of choice for handling missing data in randomized trials?. Statistical methods in medical research, 27(9), 2610-2626.

[35] Allison, Paul D. Missing Data. Quantitative Applications in the Social Sciences, nr. 136. Thousand Oaks: Sage.

If the missing data pattern appears to be MCAR, we will estimate equation (1) as stated, without any further accounting for the missing data, as the results will be unbiased.

It is difficult to show in practice if data are MAR or MNAR given the very data that are missing are needed for this[36]. Therefore, if missing data are not MCAR (based on the tests above, and could be either MAR or MNAR) we propose to run a pattern mixture model. This approach models for the observed and unobserved portion of the missing data[37] and can be undertaken in Stata by first using the `rctmiss`[38] package which can model data and missingness jointly with a pattern mixture model whereby the differences between the missing and observed data are modelled at the same time as the main effect is estimated.

Baseline missing data

In the baseline assessment data we have already observed a missing data case in relation to the age-standardised PiRA score: there are pupils included in the trial, which had sat the baseline assessment but obtained a raw score of 0 on the test. In consultation with the delivery team it was ascertained that these pupils were tested but did not provide sufficient answers for them to accrue any raw score other than 0 on the PiRA test. Because the lowest score that can be age-standardised using the PiRA manual is 1, these pupils therefore have missing baseline assessment data. There are 11 such observations (highlighted in the baseline equivalence table above). Our preferred approach, to avoid removing these pupils from the analysis is to undertake what amounts to closest-value data imputation (data replacement), and assign a raw score of 1 on this test, the next possible raw score. This would allow for the calculation of an age-standardised PiRA score and the inclusion of the 11 observations in the final analysis (providing outcome data is forthcoming for these pupils).

Compliance

To enable a non-compliance analysis, compliance will be defined at the school level, based on completion of programme activities, as recorded by the Delivery team.

We have defined “compliance” as the fulfilment of a set of minimum criteria which determine whether a school has delivered PALS-UK. This is a binary measure, indicating whether a school is compliant or not. For a measure of fidelity, i.e. the quality of implementation (above or below the compliance threshold, see below). The criteria required for a school to be deemed compliant have been defined in collaboration with the Delivery team, and consist of:

Compliance criterion	Data source	Compliance indicator
Attendance at all PALS-UK initial training session	Register of attendance	Attendance at Part 1 teacher training: at least 1 member of staff per schools attends.
Completion of the four weeks of training to the manual	Monitoring logs/Survey	Delivery of the four weeks of training: all weeks delivered in order for school to be deemed compliant.
Completion of PALS-UK delivery	Monitoring logs/Survey	Delivery of the main intervention: a minimum of 12 consecutive weeks delivered.

For a school to be deemed compliant, **all** three compliance criteria must be fulfilled. In other words, any school that only meets two of the three criteria will not be deemed fully compliant.

In a situation of imperfect compliance, whereby not all participating schools are deemed compliant using the three criteria, we will undertake a CACE (complier average causal effect) analysis, by drawing on an instrumental variable (IV) approach, and using a two-stage least squares (2SLS) estimation

[36] Fielding, S., Ogbuagu, A., Sivasubramaniam, S., MacLennan, G., & Ramsay, C. R. (2016). Reporting and dealing with missing quality of life data in RCTs: has the picture changed in the last decade?. *Quality of Life Research*, *25*(12), 2977-2983.
[37] Carpenter, J., & Kenward, M. (2012). *Multiple imputation and its application*. John Wiley & Sons.
[38] White, I. (2018). RCTMISS: Stata module to analyse a randomised controlled trial (RCT) allowing for informatively missing outcome data. https://ideas.repec.org/c/boc/bocode/s458304.html

approach to recover the treatment effect for those who complied with assignment. The first stage estimates if the assignment to PALS-UK pushes schools to take up treatment (the first stage regresses treatment assignment on compliance (as defined above). This will estimate a compliance rate. Results for the first stage will report the correlation between the instrument and the endogenous variable; and an F test. The second stage of the IV estimation predicts the outcome using the compliance rate estimated in the first regression by substituting the treatment indicator (PALSUK in Equation (1)) with the compliance rate. [39] [40] The results of this model will answer the research question: to what extent does *compliance* with PALS-UK implementation requirements lead to improved outcomes for children? This model will be estimated for the primary outcome measure only.

Intra-cluster correlations (ICCs)

The ICC is a key parameter for clustered trials. It represents the proportion of variance in a given outcome that can be explained by the variation between clusters (here: schools) as opposed to within-clusters.

One concern in using ICC estimates based on similar language measures is whether they are appropriate for the planned programme. If an inaccurate estimate for the ICC is used, the resulting sample size estimate may be either too large or too small.

The ICCs used for the power calculations reported above is based on previous EEF trials and a conservative estimate of the between-school variance.

In the final report we will report ICCs as at protocol stage (the one above); and at analysis stage. The ICC at analysis stage will be based on the primary outcome measure; and will be calculated using a model similar to Equation (1) but with no predictors, accounting for the clustering of pupils in schools (the so-called empty model).

Effect size calculation

We will use the effect sizes for cluster-randomised trials given in the EEF evaluator guidance, as adapted from Hedges:[41]

$$ES = \frac{(\bar{Y}_T - \bar{Y}_C)_{adjusted}}{\sqrt{\sigma_S^2 + \sigma_{error}^2}}$$

Where $(\bar{Y}_T - \bar{Y}_C)_{adjusted}$ is the mean difference between intervention groups adjusted for baseline characteristics and $\sqrt{\sigma_S^2 + \sigma_{error}^2}$ is an estimate of the population standard deviation (variance).

From the primary outcome model, we will take each group's mean and variance to calculate the effect size. This variance will be the total variance (across both pupil and school levels, without any covariates, as emerging from a 'null' or 'empty' multi-level model with no predictors). The ES therefore represents the proportion of the population standard deviation attributable to the intervention. [42] A 95% CI for the ES, that takes into account the clustering of pupils in schools, will also be reported. Effect sizes will be calculated for each of the models estimated.

[39] Angrist, J. D., & Keueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings?. *The Quarterly Journal of Economics*, *106*(4), 979-1014.
[40] Angrist, J. D. (2006). Instrumental variables methods in experimental criminological research: what, why and how. *Journal of Experimental Criminology*, *2*(1), 23-44.
[41] Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. Journal of Educational and Behavioral Statistics 32, 4:. 341 - 370 https://doi.org/10.3102/1076998606298043
[42] Hutchison, D., & Styles, B. (2010). *A guide to running randomised controlled trials for educational researchers*. Slough: NFER.