

Using the DreamBox Reading Plus adaptive reading intervention to improve reading attainment: A two-armed cluster randomised controlled trial

Statistical Analysis Plan



Education
Endowment
Foundation

Evaluator (institution): Manchester Metropolitan University

Principal investigator(s): Steph Ainsworth and Stephen Morris

PROJECT TITLE	Using the DreamBox Reading Plus adaptive reading intervention to improve reading attainment: A two-armed cluster randomised controlled trial
DEVELOPER (INSTITUTION)	Reading Solutions UK
EVALUATOR (INSTITUTION)	Manchester Metropolitan University
PRINCIPAL INVESTIGATOR(S)	Steph Ainsworth and Stephen Morris
PROTOCOL AUTHOR(S)	Stephen Morris and Sandor Gellen
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, Year 5, KS2
NUMBER OF SCHOOLS	123 schools
NUMBER OF PUPILS	5,404 pupils
PRIMARY OUTCOME MEASURE AND SOURCE	Reading attainment (New PiRA Summer 5 Test)
SECONDARY OUTCOME MEASURE AND SOURCE	Reading comprehension (KTEA-3 comprehension subscale) Reading fluency (KTEA-3 Silent Reading fluency subscale) Reading vocabulary (KTEA-3 vocabulary subscale) Reading self-efficacy (Feeling about Reading self-efficacy subscale) Reading motivation (Feeling about Reading motivation subscale)

SAP version history

VERSION	DATE	REASON FOR REVISION
1.2 [<i>latest</i>]		
1.1		
1.0 [<i>original</i>]	31/01/2025	N/A

Table of contents

2

3

4

4

6

7

8

10

13

13

14

16

16

19

19

21

22

23

23

25

Introduction

This statistical analysis plan describes the proposed analysis of data from a cluster randomised controlled trial (CRCT) designed to evaluate the efficacy of the DreamBox Reading Plus adaptive literacy programme.

DreamBox Reading Plus (known as Reading Plus hereafter) is an online (EdTech) adaptive silent reading programme designed to improve reading and language comprehension skills (Spichtig, A. N. et al., 2019). The programme supports fluency, comprehension (e.g. inference) and vocabulary growth, and is designed to support all readers including pupils with EAL, SEND, as well as the most able. The programme incorporates a visual skills element which scaffolds pupils' reading via a Guided Window text display to support eye movement control (Radach & Kennedy, 2013). The programme also includes additional visual skills activities to support struggling readers.

Another feature of the programme is that pupils self-select reading tasks based on age-appropriate texts. These tasks are designed to be of 'high interest' to all children, including lower attaining readers. Overall, the programme claims to have an important socio-emotional impact by building stamina and motivation to read. In addition to promoting reading proficiency directly (by improving pupils' fluency, vocabulary and reading comprehension), the programme also promises to have an indirect effect on reading attainment, through improvements in teachers' knowledge of the role of silent reading fluency in developing reading stamina and comprehension.

Year 5 teachers in intervention schools are delivering Reading Plus to whole classes for at least 90 minutes a week, over a minimum of two sessions e.g., two 45-minute sessions or three 30-minute sessions per week, for the duration of the intervention period (October 2024 to May 2025). Reading Plus is designed to supplement the school's existing reading strategy rather than replace all reading activities. It can replace some weekly guided reading lessons due to its focus on silent, independent reading, though this decision is up to the school and dependent upon their needs. While Reading Plus may replace other reading or literacy activities, this aspect will be further explored in the implementation and process evaluation. The intervention will mainly take place online within the Reading Plus platform, apart from some optional supplementary offline activities that teachers can use with pupils who might require further support. Schools may continue to use Reading Plus for the rest of the academic year if they wish.

Further details of the intervention including its theory of change can be found in the trial protocol (Gellen, et al., 2024).

Design overview

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

PRIMARY RESEARCH QUESTION

1. What is the difference in average reading attainment among Year 5 pupils in schools allocated to receive Reading Plus, 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 silent reading fluency among Year 5 pupils in schools allocated to receive Reading Plus, compared to Year 5 pupils in control schools exposed to business-as-usual conditions?
2. What is the difference in the average score for vocabulary among Year 5 pupils in schools allocated to receive Reading Plus, 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 allocated to receive Reading Plus, 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 allocated to receive Reading Plus, 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 to read among Year 5 pupils in schools allocated to receive Reading Plus, compared to Year 5 pupils in control schools exposed to business-as-usual conditions?

This is a pragmatic two-arm parallel cluster randomised controlled trial (CRCT) with whole schools allocated at random to intervention and control conditions on a 1:1 basis. The intervention is delivered to participating state primary, junior, middle or all through schools across England. The study population comprises pupils in trial schools entering Year 5 at September 2024. The primary outcome is the unstandardised score obtained by pupils on the New PIRA Summer Year 5 reading test to be administered in the summer of 2025 (score range 0-45). PiRA tests will be delivered online by schools with support from AlphaPlus Consultancy, who are subcontracted by the evaluators.

Secondary outcomes for pupils are:

- Reading comprehension subscale, Kaufman Test of Educational Achievement (Third Edition) (KTEA-3) (Kaufman & Kaufman 2014), score range 0-105
- Silent reading fluency subscale, KTEA-3, score range 0-110
- Reading vocabulary subscale, KTEA-3, score range 0-58
- Reading self-efficacy subscale, Feelings about Reading (FAR) (Carroll & Fox, 2017), score range 20-140
- Reading Motivation subscale, FAR (Vardy et al., under review), score range 10-70

The primary reading attainment outcome measure (New PIRA Summer 5 Test) and the secondary outcome measures reading self-efficacy and motivation (Feeling about Reading questionnaire) will be collected from all Year 5 pupils within participating schools. The remaining secondary outcome measures derived from the KTEA-3 assessment (discussed below) will be collected from a subsample of pupils selected at random, prior to randomisation, from within each school.

The effects of the intervention on the primary outcome will be estimated for two subgroups: 1) pupils ever-FSM, and (2) pupils designated SEND. The main trial design elements are summarised in Table 1.

Table 1. Trial design

Trial design, including number of arms		Two-armed cluster randomised controlled trial
Unit of randomisation		School
Stratification variables (if applicable)		N/A
Primary outcome	variable	Reading Attainment
	measure (instrument, scale, source)	Reading attainment raw score, 0-45, New PIRA Summer 5 Test
Secondary outcome(s)	variable(s)	Reading comprehension, Reading fluency Reading vocabulary Reading self-efficacy Reading motivation
	measure(s) (instrument, scale, source)	Reading comprehension subscale, Kaufman Test of Educational Achievement (Third Edition) (KTEA-3) 4, 0-105 Silent reading fluency subscale, KTEA-3, 0-110 Reading vocabulary subscale, KTEA-3, 0-58 Reading self-efficacy subscale, 20-140, Feelings about Reading (FAR) Reading Motivation subscale, 10-70, FAR
Baseline for primary outcome	variable	Reading attainment
	measure (instrument, scale, source)	KS1 reading teacher assessment (obtained from the NPD SRS)
Baseline for secondary outcome	variable	Reading attainment Reading self-efficacy Reading motivation
	measure (instrument, scale, source)	KS1 reading teacher assessment (obtained from the NPD SRS) Reading self-efficacy subscale, 20-140, FAR Reading motivation subscale, 10-70, FAR

As will be explained below, sample estimates of average effects will be obtained from separate regression models for each primary and secondary outcome (an adjusted analysis), where the outcome will be the dependent variable. Sample estimates of intervention effects on the primary outcome will be adjusted through the inclusion of month of birth and prior attainment in reading at KS1 as covariates in the regression model. For the comprehension, fluency, and vocabulary secondary outcomes, estimated treatment effects will be adjusted through the inclusion of month of birth and prior attainment in reading as covariates. For reading self-efficacy and motivation, the effects will be adjusted by including the baseline measures of these variables as covariates.

SELECTION OF SUBSAMPLES FOR ENDLINE TESTING – KTEA-3 ITEMS

For valid administration of the KTEA-3 instrument, trained assessors are required. To limit costs and minimise administrative burden, a sub-sample of pupils was selected at random prior to randomisation to undertake a KTEA-3 assessment at endline..

The process of selecting the pupils sampled for endline testing differed in schools depending on whether they were single or multi-form entry. For multi-form entry schools, each class was assigned a random number to seven decimal places from a uniform distribution. The class with the lowest random number was selected for sub-sampling. Within the selected class, a list of pupils was obtained, and each pupil assigned a random number to seven decimal places from a uniform distribution. Pupils were then re-ordered by these random numbers in ascending order with the top 15 pupils selected for endline testing. For single form entry schools, pupil lists were obtained, and pupils assigned a random number as previously described. Lists were arranged by these random numbers in ascending order, and the top 15 pupils sampled for endline testing.

Sampling as described took place in 122 of the 124 schools randomised. For two schools, the pupil sample for endline testing had to be drawn after randomisation because schools were late in providing the necessary list of pupils. Selecting the pupil sample after randomisation in two schools could introduce bias, as the selection might be influenced by knowledge of group assignments. However, to ensure transparency, STATA codes with a random number seed will be provided, allowing the sample selection process to remain clear and replicable

At endline, test administrators will be passed the list of sampled pupils for each school. Pupils will appear on these lists in the order determined by their random number. Test administrators will be asked to obtain 10 assessments from each school by working down the list in strict order until 10 assessments are complete. This process should mean that test administrators can complete 10 assessments in a single visit allowing for some sampled pupils to be absent on the day the test. This process also avoids arbitrary and potentially biased selection of pupils in circumstances where on the day of visit 10 sampled pupils are not all present.

MEASURES COLLECTED AT BASELINE

As discussed below, estimates of the effects of Reading Plus will come from multiple regression models that account for the hierarchical structure of the data (pupils nested in classes, and classes in schools) and will include a covariate controlling for either baseline attainment or reading self-efficacy/motivation, depending on the outcome considered. In the primary analysis the outcome is a continuous score obtained from the summer term Year 5 PiRA instrument administered at endline. A baseline measure of reading attainment, which will be entered into the primary analysis model as a covariate, will come from pupils KS1 reading teacher assessment.

Initially, schools were asked to provide KS1 reading raw and scaled scores as well as teacher assessed grade for reading at KS1 as part of the sample enumeration process. It was proposed that KS1 scaled score was preferred as a covariate in the impact analysis, as it offers finer distinctions and greater precision in measurement, as well as providing greater statistical power compared to the categorical variable of teacher assessment. However, we also considered the possibility of collecting KS1 teacher assessed grades from the National Pupil Database (NPD), in case of significant missingness in the data obtained directly from schools. Table 2 below examines the completeness of the data obtained on these three measures at baseline directly from schools:

Table 2: Baseline response – prior reading attainment, reading self-efficacy and reading motivation

Measure	N= (% of as randomised pupils)
Obtained from school records:	
KS1 Reading Raw Score	3,441 (64%)
KS1 Reading Scaled Score	3,255 (60%)
KS1 Reading Teacher Assessment	4,635 (86%)
Obtained from the Feelings About Reading questionnaire:	
Reading self-efficacy	4,795

	(89%)
Reading motivation	4,795
	(89%)
Total N=	5,404

As shown in Table 1, at the time of drafting this plan, we received valid KS1 scaled scores for only 60% of the total sample, and 20 schools did not provide any scaled scores. Conversely, we received valid reading teacher assessment data for 4,635 pupils out of a total sample of 5,404. All but one school (122 out of 123 schools who submitted pupil records) provided at least one pupil's KS1 teacher assessment. On average, each school contributed data for 44 pupils, with approximately six pupils with missing assessments per school. The amount of missing data for KS1 teacher assessments is significantly lower than that for scaled or raw scores, yet it remains higher than ideal at 14%. The primary reason for missing data in the teacher assessment records is pupils not being present in the school at the time of the KS1 assessment (n=664) in Summer 2022. Excluding pupil observations with missing KS1 prior attainment data from the primary estimation sample is unlikely to introduce bias; however, the reduced sample size will lower statistical power. As a result, it has been decided to acquire teacher assessments from the ONS SRS and not to use the measures we obtained direct from schools. In this way we hope to limit the amount of missing information in the sample at analysis and in so doing boost statistical power.

For secondary outcomes derived from the KTEA-3 research instrument, estimates of the intervention effect will also come from multiple regression models containing a baseline measure of reading attainment in the form of the reading score at KS1 as a covariate. For the secondary outcomes reading self-efficacy and reading motivation, these measures come from the Feelings About Reading questionnaire which was administered online to pupils at baseline – during June and July 2024 – and which will be administered to the sample again at endline. For regression models with self-efficacy and motivation outcomes, the baseline measure on the relevant outcome will be included as a covariate.

RANDOMISATION

Over 14,000 schools were approached by the Delivery Team via email marketing, social media, and networking. 392 schools completed an Expression of Interest form (either via the EEF website or directly with Reading Solutions). Of these, 130 signed the MOU, and 6 schools subsequently withdrew before randomisation. Randomisation took place on the Friday 12th July, 2024. In total 124 schools were randomised 1:1 to intervention and control groups. The Delivery Team were informed about the outcome of randomisation on the same day, and schools notified shortly after.

Initially, the plan was to achieve balance on key school-level covariates by stratifying randomisation based on average prior school performance in KS2 reading tests from the last three available summer assessments, dividing schools into terciles, and categorising them as either single or multi-form entry. However, due to an error in the code used to allocate schools (an earlier version of the baseline dataset was mistakenly linked to the randomisation outcome dataset), randomisation into intervention and control groups was conducted without stratification. As a result, complete randomisation was implemented instead of the intended stratified randomisation. This error was identified only after schools had been informed of their group allocations, making it impossible to repeat the process. An amended protocol will outline the implications of this change for the trial design and analysis.

Overall, the number of schools remained well balanced between the treatment and control groups, and the randomisation process largely achieved a good balance across each stratum, despite the strata not being used in the randomisation. Due to this error, a slight imbalance is noted in two of the six strata (see Table 3). However, as the randomisation was still governed by a random process, the overall integrity of the process was preserved. Further details on the potential impact of baseline imbalance for other relevant variables will be addressed in the corresponding section of the SAP.

Table 3. Randomisation allocation across strata (class-form entry, school-level KS2 scores)

	N Schools	N Pupils	Higher tercile KS2 score		Middle tercile KS2 score		Lower tercile KS2 score	
			Single form	Multi- form	Single form	Multi- form	Single form	Multi- form
Intervention	62	2,748 ¹	15	6	7	13	8	13
Control	62	2,656	15	6	9	12	9	11
Total	124	5,404	30	12	16	25	17	24

Randomisation was performed in STATA v18 using the command `randtreat` (Carril, 2017).

Sample size calculations overview

At the protocol stage sample size calculations were presented assuming: 1) no attrition of schools from the sample by endline, 2) five percent and 3) 10 percent attrition (Gellen, et al., 2024). These calculations are replicated in Table 3 below along with additional calculations based on the sample sizes achieved at randomisation.

Sample size calculations are based on a range of assumptions as well as judgements as to the adequacy of different sample sizes based on associated minimum detectable effect sizes (MDESs). The MDESs are assessed to determine how far they seem plausible given results from similar studies. To arrive at a plausible MDES and therefore sample size, we considered results from other studies of Reading Plus as well as studies of other similar reading programmes targeting primary school children. A randomised trial of Reading Plus in the US in the fifth grade (equivalent to Year 6), yielded effect size $ES = 0.18$ ($p < 0.001$) (Spichtig, A. N. et al., 2019). An evaluation of Lexia Reading Core, in England, but with younger pupils, obtained an effect size on the primary outcome of $ES = 0.08$ ($p = 0.15$) for all pupils, and $ES = 0.18$ ($p = 0.04$) for FSM pupils (Tracey, L et al., 2022). There are a number of US studies of Peer Assisted Learning Strategies (PALS) that found quite large effect sizes, ranging from 0.23 to 0.71, but samples contained only small numbers of fifth grade pupils and focused on students with learning disabilities (U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse, 2012). The Reciprocal Reading programme looked at reading attainment in KS2 (O'Hare, L & et al., 2019). The authors obtained $ES = 0.14$ ($p < 0.01$) and $ES = 0.18$ ($p < 0.001$) for reading comprehension.

¹ Figures relating to pupils reflect the 123 schools that have submitted pupil records at the time of writing the SAP.

Table 4. Minimum detectable effect size for reading attainment

		Protocol						Randomisation					
		No attrition		5% attrition		10% attrition		No attrition		5% attrition		10% attrition	
		OVERALL	FSM	OVERALL	FSM	OVERALL	FSM	OVERALL	FSM	OVERALL	FSM	OVERALL	FSM
Minimum Detectable Effect Size (MDES)		0.18	0.20	0.18	0.20	0.19	0.21	0.18	0.19	0.18	0.20	0.19	0.20
Pre-test/ post-test correlations	level 1 (pupil)	0.36	0.36	0.36	0.36	0.36	0.36	0.36	0.36	0.36	0.36	0.36	0.36
	level 2 (class)	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
	level 3 (school)	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
Intraclass correlations (ICCs)	level 2 (class)	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
	level 3 (school)	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
Alpha		0.05	0.05	0.05	0.05	0.05	0.05	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	0.8	0.8	0.8	0.8	0.8	0.8
One-sided or two-sided?		Two	Two	Two	Two	Two	Two	Two	Two	Two	Two	Two	Two
Average cluster size – level 1 per level 2		30	8	30	8	30	8	27	10	27	10	27	10
Average cluster size – level 2 per level 3		1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7	1.7
Number of schools	intervention	63	63	60	60	57	57	62	62	59	59	56	56
	control	63	63	60	60	57	57	62	62	59	59	56	56
	total	126	126	120	120	114	114	124	124	118	118	112	112
Number of pupils	intervention	3,213	857	3,060	816	2,907	775	2,748 ²	888	2,611	844	2,473	799
	control	3,213	857	3,060	816	2,907	775	2,656	849	2,523	807	2,390	764

² Number of pupils are based on information provided at the at the time of drafting the SAP: 123 out of 124 submitted data, some with missing FSM status

	Protocol						Randomisation					
	No attrition		5% attrition		10% attrition		No attrition		5% attrition		10% attrition	
	OVERALL	FSM	OVERALL	FSM	OVERALL	FSM	OVERALL	FSM	OVERALL	FSM	OVERALL	FSM
total	6,426	1,714	6,120	1,632	5,814	1,550	5404	1,737	5,134	1,650	4,863	1,563

Together, these are the most relevant studies for which high quality evidence of effect sizes are available and based on this we judge that our trial sample will need to be capable of detecting an effect size of around $ES=0.18$ with 80 per cent power. According to (Kraft, 2020 Table 2, p. 250) this would rank as a medium sized effect in the context of contemporary field trials in education.

The MDESs in the table above are calculated using the PowerUp in Excel (Dong & Maynard, 2013).

The following assumptions and choices are reflected in our calculations:

- Statistical tests of the nil null hypothesis will be two-tailed tests and performed at the 95 per cent level of statistical significance with statistical power of 80 per cent.
- Randomisation of schools was carried out 1:1 to intervention and control condition
- Proportion of variance explained in the outcome by Level 1 covariates of 0.36. We do not have reliable evidence on the correlation between KS1 reading attainment (the main pre-test covariate) and PiRA Year 5 summer test (the primary outcome). We do know that the correlation between PiRA and KS2 reading attainment is about 0.73 (Lewin, C et al., forthcoming 2024). Thus, we assume it is lower for KS1. We also allow for some moderate improvement in the outcome variance explained at the class and school levels, assuming that the regression model from which impact estimates will be obtained, containing a covariate capturing prior attainment, will reduce variance explained at the pupil level but also school and class levels.
- Based on evidence from the recently completed PALS-UK trial, we assume intra-cluster correlations (ICC) of 0.10 at the school level and 0.04 at the class level (Lewin, et al, forthcoming 2024)

Given the trial's relatively short duration and the Delivery Team's confidence in retaining schools within the study we initially set a target of recruiting 126 schools. Based on sample sizes obtained in the recently completed PALS-UK trial, we assumed that on average there would 1.7 classes per school with around 30 pupils in each class (Lewin, C et al., forthcoming 2024). These sample size assumptions, together with the assumptions set out at the bullet points above led to MDESs at the protocol stage of 0.18 (with either zero or five per cent school attrition) and 0.19 (with 10 per cent school attrition) (Gellen, S et al., 2024).

As shown in the table above, 124 schools were randomised into intervention and control groups, with an average of 1.7 classes per school and 27 pupils per class. Based on these figures, the MDESs calculated at the randomisation stage were similar to those established during the protocol phase: 0.18 with no attrition, 0.18 with 5% attrition, and 0.19 with 10% attrition.

For the FSM subgroup, the protocol phase estimated MDESs of 0.20 with no attrition, 0.20 with 5% attrition, and 0.21 with 10% attrition. At the randomisation stage, however, the observed proportion of FSM-eligible pupils was higher than expected (with an average of 10 FSM-eligible pupils per class). Consequently, the MDESs for this subgroup were slightly lower: 0.19 with no attrition and 0.20 for both 5% and 10% attrition.

Analysis

The analysis will proceed based on the principle of intention to treat (ITT). That is, pupils are identified in the analysis as members of the intervention or control group based on their school's allocation to intervention and control conditions at randomisation, regardless of whether the school subsequently complies with their experimental group status. 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.

PRIMARY OUTCOME ANALYSIS

The primary analysis seeks an estimate of the average effect of intention to treat (AITT), for the intervention, on reading attainment. The measure of reading attainment is derived from the new PiRA summer term Year 5 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, as well as a having a high correlation with external measures of attainment (McCarty & Ruttle, 2018). The PiRA Year 5 summer test, comprises 30 items with a scoring rubric giving rise to a minimum score of zero and maximum of 45.

A sample estimate of the average effect of intention to treat, for Reading Plus on reading attainment will be obtained from a linear mixed regression model as follows:

$$y_{ijk} = \beta_0 + \beta_1 t_k + \beta_2 x_{ijk} + \beta_3 a_{ijk} + v_k + v_{jk} + \varepsilon_{ijk} \dots \dots \dots [1]$$

Here y_{ijk} is the unstandardised raw reading score obtained by pupil i in class j and school k . The variable t_k takes the value one if school k is allocated to the intervention, zero otherwise. The covariate x_{ijk} is the pupil's teacher assessed reading grade obtained in their KS1 reading test and a_{ijk} a covariate capturing age in months.

The terms β_0 to β_3 are the unknown model parameters that will be estimated from the data. The sample estimate of the parameter β_1 is interpreted as the adjusted sample average effect of the programme on reading scores. Uncertainty in model estimates will be evaluated based on 95% frequentist confidence intervals as well as continuous p-values for tests of the nil null hypothesis.

The model takes account of pupils being clustered or nested in classes, and classes in schools. There are three random effects in the model. The random effect ε_{ijk} is a pupil level random effect, whilst v_{jk} and v_k are class and school level random effects respectively. Each random effect is assumed to be normally distributed with zero mean in the population and with variances σ_k^2 , σ_j^2 and σ_ε^2 respectively. The random effects are also assumed to be uncorrelated with each other and uncorrelated with the covariates in the model. The proportion of the total outcome variance at the school and class levels can be expressed as intraclass correlation coefficients, and written as follows, first for the class level:

$$\rho_j = \frac{\sigma_j^2}{\sigma_k^2 + \sigma_j^2 + \sigma_\varepsilon^2}$$

And for the school level:

$$\rho_k = \frac{\sigma_k^2}{\sigma_k^2 + \sigma_j^2 + \sigma_\varepsilon^2}$$

The primary analysis model will be estimated using the statistical software STATA v18, and the command 'mixed' on the basis of restricted maximum likelihood. Sample estimates of both ρ_j and ρ_k and their respective confidence intervals can be calculated directly from the STATA v18 model output or through the command `estat icc`. Sensitivity analysis and further statistical models relating to the primary analysis are discussed in the 'additional analyses' section. The sample estimate of the intervention effect will be converted into an effect size (for more detail, see the 'effect size calculation' section).

SECONDARY OUTCOME ANALYSIS

Secondary analysis will involve estimating the effects of Reading Plus on secondary outcomes from two sources: the KTEA-3 assessment instrument and the Feeling about Reading (FAR) questionnaire.

Turning first to secondary outcomes from the KTEA-3 assessment, three outcomes will be considered:

- Silent reading fluency
- Comprehension, and
- Vocabulary.

KTEA-3 comprises a series of measures that have been shown to be reliable and valid (Breux & Lichtenberger, 2016). The three KTEA-3 subtests that will be administered in this study take approximately 30 minutes to complete in total. The reading comprehension subtest gives possible raw scores of 1 to 31, which convert to weighted scores of 0 to 105. Weighted scores need to be calculated for this subtest to account for the fact that the subtest is adaptive and therefore children will not all complete the same items. The reading vocabulary subtest gives possible raw scores of 0 to 58 and the silent reading fluency has a range of 0 to 110. While the reading vocabulary subtest is also adaptive (children can 'reverse' to easier items if they score 0 in the first 3 items), a weighted score is not needed because the scoring gives the child credit for items earlier than the basal level (i.e. items which did not need to be administered because they would have been too easy). A weighted score is also not required for the silent reading fluency subset because for this subtest all children are given the same items and are asked to read as many of them as they can (and identify whether each statement is true or false) within two minutes.

As discussed above, these assessments will be conducted with 10 students per school selected at random prior to randomisation. The KTEA-3 measure will be administered by trained test assessors from AlphaPlus Consultancy.

In multiform entry schools – as with single form entry schools – all pupils that complete a KTEA-3 assessment will be from the same class. Therefore, the sample will be clustered at the school/class level only. Sample estimates of the effect of Reading Plus on pupils for these three outcomes will be obtained from a model of the following form, where the usual assumptions are made:

$$y_{ij} = \beta_0 + \beta_1 t_j + \beta_2 x_{ij} + \beta_3 a_{ij} + v_j + \varepsilon_{ij} \dots \dots \dots [2]$$

Here the variable y_{ij} is the response – either pupil i 's silent fluency, comprehension or vocabulary score in school j - depending on which of the three KTEA-3 secondary outcomes is being considered. x_{ij} the KS1 teacher assessed reading score for pupil i in school j . The covariate a_{ij} is as before with v_j a school-level random effect. These models will be estimated in STATA v18 using the `mixed` command and restricted maximum likelihood. Estimates of the intraclass correlations for each model will be obtained directly from the STATA v18 model output or using the command `'estat icc'`.

Two secondary outcome measures will be obtained from the FAR:

- Reading self-efficacy
- Reading motivation

The motivation to read scale developed by Vardy et al. (in prep) has been shown to have high reliability (Cronbach's $\alpha = 0.83$). The reading self-efficacy scale is adapted from Carroll & Fox's (2017) original version of the scale with minor revisions to the phrasing of a few items. This has a Cronbach's α value of 0.90 (Vardy et al., in prep). The reading self-efficacy scale is comprised of 20 items with responses on a seven-point scale. The minimum score available is 20 points the maximum 140. The reading motivation score is derived from 10 items with responses also on a seven-point scale, leading to a minimum possible score of 10 and maximum of 70.

Sample estimates for the effect of Reading Plus on both motivation and self-efficacy will be obtained from estimating two regression models (one for each outcome) similar to that used in the primary analysis, except where y_{ijk} is either the self-efficacy score for pupil i in class j and school k or similarly a pupil's motivation score – depending on which of the two models is considered. In these models x_{ijk} is the baseline score for either self-efficacy or motivation depending on the response.

These models will also be fitted in STATA v18 statistical software using the command `'mixed'` and restricted maximum likelihood. Estimates of the intraclass correlations for each model will be obtained directly from the STATA v18 model output or using the postestimation command `'estat icc'`.

SUBGROUP ANALYSES

Subgroup analysis will be performed for two subgroups: those ever-FSM and pupils with a SEND designation. Variables identifying these pupils in our sample have been obtained directly from schools, and except for two schools, prior to randomisation.

Subgroup analysis will involve fitting two models for each of the two subgroups to be considered. The first model for the subgroup in question will have the same form as the primary analysis regression model but will be estimated on a sample restricted to pupils that are members of the group. Results from this model will show whether pupils' ever-FSM or pupils' SEND benefit from the intervention relative to their counterparts in the control group.

The second model for each subgroup will be estimated on the full sample and will contain an interaction between the subgroup indicator and the intervention group indicator, as follows:

$$y_{ijk} = \beta_0 + \beta_1 t_k + \beta_2 g_{ijk} + \beta_3 t_k g_{itk} + \beta_4 x_{ijk} + \beta_5 a_{ijk} + v_k + v_{jk} + \varepsilon_{ijk} \dots \dots \dots [3]$$

The terms in the model are as before with the exception of g_{ijk} , which is coded to '1' if pupil i in class j and school k is ever-FSM or SEND, depending on which group is the focus of the analysis. The sample estimate of β_3 is interpreted as the extent to which, members of whichever group is being considered, have benefited disproportionately from their participation in Reading Plus.

Models will be fitted on the data in the statistical package STATA v18 using the command 'mixed', and restricted maximum likelihood. For all models we will calculate the intraclass correlation coefficients at the school and class levels using the STATA postestimation command `estat icc`.

ADDITIONAL ANALYSES

Two sets of additional analysis are proposed. In the first, we will estimate a series of further models to support and further expand on the primary impact analysis described above. In the second, we proposed to undertake a mediation analysis exploring the extent to which gains in fluency among pupils might mediate the causal effect of Reading Plus on reading attainment. This latter analysis will be undertaken if there is evidence that intervention has had a meaningful effect on the primary outcome and there is sufficient data.

Additional analyses relating to the primary outcome model

Analyses in support of the main primary analysis will comprise sensitivity testing the main primary analysis regression as well as fitting an additional linear outcome model using ordinary least squares (OLS) regression.

The linear OLS model will produce sample estimates of the effect Reading Plus on reading attainment with a slightly different interpretation to the primary model. The benefit of this model is that it rests on fewer distributional assumptions when compared to the primary analytical approach. Sample estimates of the effect of the intervention on reading in the primary analysis are obtained from a linear mixed model, fitted using restricted maximum likelihood, that accounts directly for clustering of pupils in classes and classes within schools, using random effects. Model estimates are therefore interpreted as cluster specific effects and the variances at the various levels of clustering are estimated directly. Alternatively, it is legitimate to model the relationship between the intervention and primary outcome by averaging over the levels in the data and simply adjusting standard errors to take account of clustering. Such a model can be estimated using ordinary least squares linear regression and cluster robust standard errors in STATA v18. With samples of the size anticipated, these two types of causal effect estimates should be quite similar to one another. As explained in more detail below, the proposed statistical models for our missing data, mediation and compliance analyses will be run in STATA, which does not support hierarchical structures for these procedures but does allow for the use of cluster robust standard errors. Therefore, the model fitted using OLS linear regression will act as a benchmark for these analyses.

Turning first to sensitivity analysis. We propose to fit the following additional models in support of the primary analysis:

- A variance components model which enables us to assess the unconditional proportions of the outcome variance at the levels in the model and from which we will obtain the denominator to be used in effect size calculations in the primary outcome analysis³. This takes the form:

$$y_{ijk} = \beta_0 + v_k + v_{jk} + \varepsilon_{ijk} \dots \dots \dots [4]$$

- A model as above but in addition containing the intervention group indicator t_k , to provide an unadjusted estimate of the intervention effect and through comparison enable us to assess the effects on our results of covariates in the primary analysis model.
- A model of the following form containing testing for the effects of different pre-test covariates:

$$y_{ijk} = \beta_0 + \beta_1 t_k + \beta_2 x'_{ijk} + \beta_3 l_{ijk} + \beta_4 x_{ijk} + \beta_5 a_{ijk} + v_k + v_{jk} + \varepsilon_{ijk} \dots \dots \dots [5]$$

Along with the teacher assessed grade pupils obtained in their KS1 reading tests, we also attempted to collect KS1 reading scaled and raw scores prior to randomisation direct from schools for all pupils in the enumerated sample. Unfortunately, there is quite a bit of missing data on the raw and scaled continuous KS1 reading test scores. However, we do want to examine how far including continuous scaled scores at KS1 as pre-test covariates in our model might improve the precision of our sample estimates of β_1 . To do this we propose to estimate a model as [1] above with the inclusion of two additional terms. The first of these is x'_{ijk} which represent pupil i in class j and school k 's scaled score in their reading KS1 assessment. For each case where we fail to observe x'_{ijk} the missing value is replaced by zero. This is included in the model alongside the teacher assessment they received - x_{ijk} . The second additional covariate is a binary indicator l_{ijk} . This is set to '1' if the pupil has a missing value (or zero) in x'_{ijk} , zero otherwise.

- A model identical the main primary analysis model except for the inclusion of additional covariates: baseline FAR self-efficacy and motivation test scores, FSM, gender, SEND, EAL, school size (in form entry), and average school-level KS2 scores for the last three years in reading. This final model contains further covariates (except for scaled and raw KS1 reading scores) to explore whether any further gains in power can be achieved.

For each of these models, the intraclass correlation coefficients will be calculated for the school and class levels using the STATA post-estimation command `estat icc`.

As discussed above, we will fit a model equivalent to the main primary impact regression model but using linear OLS with cluster robust standard errors, taking into account heteroskedasticity and clustering at the school level and using 'cluster' version of the `hcc2` robust standard error (MacKinnon et al., 2022), often preferred for standard error estimation in the case of randomised trials⁴. For further robustness a degrees of freedom adjustment will be made (Bell & McCaffrey, 2002). The model takes the following form:

$$y_{ij} = \beta_0 + \beta_1 t_j + \beta_2 x_{ij} + \beta_3 a_{ij} + \beta_4 s_j + \varepsilon_{ij} \dots \dots \dots [6]$$

³ This is line with the EEF Statistical analysis guidance that states, 'evaluators should use the unconditional standard deviation in the calculation of ES. [...] For transparency, evaluators should provide all parameters [including all variance components] to allow third parties to compute the ES of their interest.' (2022, p.7-8)

⁴ See <https://oes.gsa.gov/assets/files/calculating-standard-errors-guidance.pdf>

Where i indexes for pupil and j school. As mentioned above this model will act as a benchmark for missing data, mediation and compliance analyses. The linear regression model is also attractive because it requires fewer assumptions to be made compared linear mixed effects (multilevel) models used in the primary and secondary analysis.

Uncertainty in our results will be represented through frequentist 95% confidence intervals and continuous p-values.

Testing reading silent fluency as a mediator

Mediation analysis will be conducted to examine ‘fluency’ as a key mediator of the intervention effect. Gains in silent reading fluency are hypothesised to be a key process or mechanism through which Reading Plus leads to improvements in reading attainment.

In counterfactual language we are interested in the natural indirect effect (NIE) (Pearl et al., 2016)⁵. In this analysis, we assume that a total effect of Reading Plus on reading attainment exists, and we focus on decomposing this effect by its direct component as well as its indirect component via fluency. Using the definition of NIE from Imai et al. (2013) we can define the NIE in our case as representing the causal effect of Reading Plus on reading attainment transmitted through changes in a student’s reading fluency following their receipt of Reading Plus⁶. We will obtain sample estimates of the NIE due to fluency, the natural direct effect (NDE) and the total effect (TE) of Reading Plus on reading attainment, using the command `mediate` in STATA v18.

As both fluency and reading attainment are continuous, and we assume normally distributed measures, the analysis can be performed within a linear framework. This analysis, as with the compliance and missing data analysis, will be conducted such that standard errors are adjusted for clustering at the school level. The approach is operationalised within the `mediate` command, and in the case of linear models, is similar to that of Baron & Kenny (1986) except for allowing an interaction between the intervention group indicator

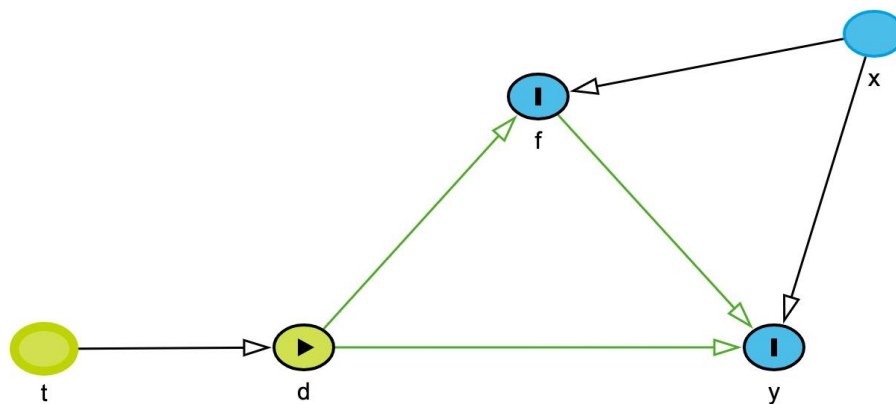
and the mediator (StataCorp LLC, 2023).

We present a simplified directed acyclical graph⁷ (or DAG) of our mediation model. The DAG sets out the relationships between the key variables in our model and encodes our assumptions about how they are causally related to one another. A causal relationship is represented by an arrow between the variables in the graph with the arrowhead representing the direction of causality:

⁵ Where the NIE in potential outcome terms is $NIE = E[Y_{ij}(1), M_{ij}(1)] - E[Y_{ij}(1) - M_{ij}(0)]$, and where $Y_{ij}(1)$ is the potential outcome for pupil i in school j under treatment and $M_{ij}(1)$ the potential fluency score for pupil i in school j under treatment with $M_{ij}(0)$ the potential fluency score for pupil i in school j under control.

⁶ This effect is sometimes referred to as the total natural indirect effect (TNIE), causal mediation effect (CME), or average indirect treatment effect (AITE) and should not be confused with the pure natural indirect effect (PNIE).

⁷ This DAG is produced using the online resource <https://www.dagitty.net/>



With reference to the primary analysis model, two additional variables are introduced: d and f . The node t represents randomisation, which determines whether a pupil is allocated to the intervention or control group, which in turn is represented by the variable d . The variable f represent a continuous silent reading fluency score obtained from the KTEA-3 assessment and which is the mediating variable in the model. As before y , the outcome, is reading attainment and x a pre-randomisation measure of reading attainment.

Our modelling strategy warrants a causal interpretation if the following assumptions hold (StataCorp LLC, 2023):

- The relationship between the reading attainment y and the intervention d is unconfounded – which we assume to be true due to randomisation;
- The relationship between fluency f and reading attainment y is confounded by prior reading attainment x which we control for in the analysis through the inclusion of prior attainment as a covariate. Will explore whether there is any evidence of other measured confounds.
- The relationship between the intervention d and the mediator f is unconfounded – which we assume is again true due to randomisation; and
- Prior reading attainment x is not caused by the intervention d . This is true by design because x relates to a period prior to d

We will in effect estimate the following statistical models to obtain an estimate of the NIE using the `mediate` command, though the algorithm produces the model estimates from a single command line:

$$y_{ij} = \beta_0 + \beta_1 d_j + \beta_2 f_{ij} + \beta_3 d_j f_{ij} + \beta_4 x_{ij} + \varepsilon_{ij} \dots \dots \dots [7]$$

$$f_{ij} = \alpha_0 + \alpha_1 d_j + v_{ij}$$

The error terms ε_{ij} and v_{ij} are assumed to be uncorrelated and to have zero mean with variances σ_ε^2 and σ_v^2 . The STATA program provides estimates of potential outcome means for values of the treatment group indicator and mediator, yielding estimates of the NIE, NDE and TE. Uncertainty captured through frequentist 95% confidence intervals. Because the fluency test score is obtained for 10 pupils from each school selected at random, the sample upon which this analysis will be conducted is smaller than that for the primary analysis and is clustered at the school level only.

LONGITUDINAL FOLLOW-UP ANALYSES

No longitudinal follow-up is proposed.

IMBALANCE AT BASELINE

We plan to compare the characteristics of schools and pupils in the intervention and control groups for both the 'as randomised' and 'as analysed' samples. The 'as randomised' sample will include all schools and pupils who did not withdraw from the study after randomisation, while the 'as analysed'

sample will consist of all pupils for whom KS1 Teacher Assessment results and PIRA Summer 5 Test scores are available. We will present tables comparing counts and proportions for categorical variables, and means and standard deviations for continuous variables, by intervention and control groups across the 'as randomised' and 'as analysed' samples by the following pupil-level variables: gender, age in months, FSM, SEND, EAL, KS1 Teacher Assessment, and baseline FAR self-efficacy and motivation scores. Similarly, cross-tabulations will examine intervention and control group differences/similarities by school-level characteristics, where variables covering school area, Ofsted ratings, and school type, and continuous measures including the proportion of FSM pupils, EAL pupils and average KS2 scores will be considered.

As of drafting the SAP, the evaluation team has received data on nearly all pupil-level characteristics. Apart from the previously mentioned missing KS1 baseline scores, we have complete data on pupil characteristics from 122 out of the 124 schools randomised, whilst one school have provided partial information and one school have yet to submit pupil records. A descriptive analysis was conducted after randomisation to compare pupil and school characteristics between the intervention and control groups, focusing on school-, and pupil-level attributes. This comparison appears particularly valuable given that no stratification was used during randomisation, as the process followed a completely randomised design.

These comparisons, summarised in the table below, indicate that randomisation produced two well-balanced groups. The sample achieved a strong balance between the treatment and control groups regarding the proportions of FSM, SEND, and EAL status, as well as prior reading attainment at both the school and pupil levels. Our judgement is that these groups are also well balanced in terms of geographical distribution and form of entry.

Table 4. Baseline balance of the 'as-randomised' sample

		Control	Treatment
School-level			
Number of schools		62	62
Form of entry	1	33	30
	2	21	20
	2 or more	8	12
Average of KS2 score		104.5	104.8
Region	East Midlands	3	6
	East of England	8	4
	London	7	6
	North East	3	2
	North West	16	18
	South East	7	8
	South West	5	2
	West Midlands	6	6
	Yorkshire and The Humber	7	10
Pupil level			
Number of pupils		2656	2748
FSM proportion	FSM	32%	32%

	non-FSM	64% ⁸	68%
SEND proportion	SEND	20.1%	21.6%
	non-SEND		
EAL proportion	EAL	26.1%	24.7%
	non-EAL	73.9%	75.3%
Average KS1 scaled score		101.1	101.1
Average KS1 raw score		24.4	25.2
KS1 Teacher Assessment	Missing	17.2%	11.4%
	BLW (Below)	3.5%	4.5%
	PK1-PK6 (Pre-Key Stage 1-6)	5.6%	4.5%
	WTS (Working Towards)	20.8%	23.0%
	EXS (Expected Standard)	40.3%	44.3%
	GDS (Greater Depth Standard)	12.6%	12.2%

MISSING DATA

For the primary analysis, we will conduct sensitivity tests to evaluate whether missing data at the endline introduces bias or leads to imprecise estimates of β_1 . Missingness before randomisation is unlikely to cause bias in intervention effect estimates, though it may lead to a decline in power due to diminished sample sizes. To ensure that missingness due to missing data in pre-randomisation covariates is minimised, we will acquire KS1 Teacher Assessments for our sample from the ONS SRS.

Post-randomisation missingness in the primary outcome could arise due to several factors, including but not limited to:

- Parents withdrawing their children from the study and requesting data deletion
- Pupils leaving the school before completing the New PiRA Summer 5 Test
- Schools withdrawing from the evaluation and asking for data deletion
- Pupils being absent on the day of testing, thus unable to provide outcome data

Initially, we will assess the nature of the missing data, determining whether it is missing completely at random (MCAR), missing at random (MAR), or potentially missing not at random (MNAR). This will include calculating and comparing missing data rates across trial arms. If missingness exceeds 5% in both control and treatment groups, we will investigate whether baseline covariates, such as gender, FSM status, SEND status, EAL status, school size, KS1 scores, and FAR measures at baseline explain the missing data by fitting a logistic regression model. Furthermore, subsequent bivariate tests will be carried out to assess the correlation between the presence of an endline New PiRA Summer 5 Test score and other explanatory variables: gender, FSM status, SEND status, EAL status, school size, baseline self-efficacy scores, and baseline motivation scores. Covariates associated with New PiRA Summer 5 Test scores or significantly related to missingness (at a 95% confidence interval) will be identified as potential explanators for the missing outcome data and used incorporated into a multiple imputation model.

If the missing data rate exceeds 5% in either trial arm, and the dropout model shows that missingness is associated with covariates, further sensitivity tests will evaluate the impact of this missing data on sample estimates using multiple imputation by chained equations (MICE). The multiple imputation

⁸ One school has not yet provided FSM information, which is why the percentages for FSM and non-FSM in the control group do not add up to 100%.

process will involve specifying an imputation model for each variable with missing data, determining the creation of 10 imputed data sets. The imputation will be performed using STATA v18, which supports only single-level imputation. Although multilevel imputation is available in R, the process is slow and limited to two levels. To be consistent with the imputation model, we will first re-estimate the primary outcome model described in the primary analysis section using ordinary least squares (OLS) with cluster-robust standard errors, adjusting degrees of freedom via the `hc2` method. The STATA command used will be `"vce(hc2 [cluster_name], dfadjust"` to account for clustering at the school level. This OLS model will provide a population average treatment effect, which we will compare with results from the imputed datasets.

The MICE procedure will generate multiple datasets with missing values imputed. We will then use `"mi estimate"` to re-estimate the intervention effect by running OLS regressions on these imputed datasets. Comparing these results with the analysis described in the previous paragraph will help determine whether the missing data follows an MAR process.

If there is some evidence that missing data are MNAR, or we cannot exclude the possibility, it will become more challenging to assess the consequences of attrition for our sample estimates. If there is a significant imbalance in attrition between treatment and control groups, we will estimate intervention effects using Lee bounds, which can be used to assess the potential impact of this imbalance. The bounds are calculated for the always-responding subsample, not the full sample, which may limit the generalisability of the findings. The Lee (2009) estimator provides bounds on treatment effects under minimal assumptions—namely, randomisation and the monotonicity assumption (i.e., no subjects are more likely to respond under control than intervention conditions). Lee bounds create trimmed bounds by removing (or trimming) observations from the group with the higher response rate (treatment group) to match the response rate of the group with the lower response rate (control group). This is done under the assumption that those in the treatment group who responded represent the same distribution as those who would have responded had they been in the control group. This creates a conservative estimate that bound the treatment effect. If justified, we will calculate these bounds using the `"leebounds"` package in STATA v18 (Tauchmann, 2014).

COMPLIANCE

Due to the nature of the intervention and the control developers can exercise over access to the Reading Plus platform, non-compliance affects the intervention group only. Control group pupils will not be able to access Reading Plus. Thus using the terminology of Gerber & Green (2012) we have one sided non-compliance.

Several potential compliance indicators were discussed with the delivery team and the EEF. These indicators range from those defined solely at the school level to those that combined school and pupil level compliance. The indicators considered are listed below. A binary compliant/non-compliant measure for each of these indicators can be derived from the Reading Plus management information system for each pupil and school. The indicators are:

- The school fails to attend training in Reading Plus – thus all schools and pupils therein assigned to the intervention group but where no representative from the school attends training and there is no further activity are considered non-compliant.
- The school attends training but there is no further evidence that the school uses the Reading Plus platform.
- The school attends training and there is evidence that Reading Plus is used in the school but there are pupils who have engaged with fewer than 15 texts over the three terms the programme runs for.
- The school attends training and there is evidence that Reading Plus is used in the school but there are pupils that engaged with 15 or fewer texts over the first two terms and no texts in the third term.

For the purposes of compliance analysis, it has been agreed with the developers and EEF, that we will use definitions 3 and 4. It was judged that engaging with fewer than 15 texts could have no influence on pupils' reading. There was some concern that pupils might engage with 15 texts in the final term of implementation just before they complete endline reading assessments, and that such minimal engagement in close proximity to assessment could possibly influence their subsequent performance in tests. Thus, this final indicator will be used as a sensitivity check.

Because the definitions of compliance that will be the focus of the analysis captures pupil level engagement with the platform, the compliance analysis will be performed with 'pupil' as the unit of analysis. Compliance analysis will be carried out using the two-stage least squares instrumental variables approach and the command `ivregress 2sls` in STATA v18. Standard errors will be clustered at the school level. The analysis first effectively involves estimating a model that predicts compliance where one of the two indicators above is used to construct a binary measure of compliance and used as the dependent variable. d_{ij} below is coded to '1' if the pupil is, according to the indicator considered, compliant, zero otherwise.

$$d_{ij} = \gamma_0 + \gamma_1 t_j + \gamma_2 x_{ij} + \gamma_3 t_j x_{ij} + \gamma_4 c_j + \gamma_5 t_j c_j + \varepsilon_{ij} \dots \dots \dots [8]$$

The model contains the variable t_j which is the school level indicator capturing whether the school was assigned to the intervention or control group, and the pre-test measure of reading x_{ij} as well as its interaction with t_j . In addition, the average reading score for the school in the last three KS2 summer reading tests c_j is also included in the model as a measure of school performance, along with its interaction with t_j . The model assumes that for pupils in intervention schools, their probability of compliance is a function of their prior attainment and historic school performance. For example, it may be that pupils with higher prior attainment in schools with higher historic test score performance are

more likely to comply. The predict probability $\hat{p}(ij = 1 \vee t_j, x_{ij}, c_{ij})$ is obtained from this regression and used in a second stage regression where reading attainment at endline is the dependent variable:

$$y_{ij} = \beta_0 + \beta_1 \hat{p} \dots \dots \dots [9]$$

The estimated value for the parameter β_1 is interpreted as an estimate of the effect of reading plus on reading attainment for those that comply, or the complier average causal effect. Because there is no non-compliance possible in the control group, this estimate can also be interpreted as the average effect of treatment on those treated. Moerbeek & Schie (2019) point out that complier average causal effects estimated in this way can be problematic when association between t_j and d_{ij} in the first stage regression is weak. We will report the results from a partial F-test for the first stage model above, where value of 10 or higher is generally considered to be acceptable.

INTRA-CLUSTER CORRELATIONS (ICCs)

The ICC associated with school and class for the primary outcome will be reported along with a 95% confidence interval. The ICC will be calculated for the primary analysis model, as well as from an empty model (i.e., one without covariates).

For more detail, please refer to the primary outcome analysis (p. 11-12) and additional analyses (p. 14-15) sections above.

EFFECT SIZE CALCULATION

For each model discussed above, except for those estimated using ordinary least squares, results from the models will be presented as effect sizes. That is, as standardised difference in means that

are consistent with Hedges g , given the relatively large samples likely at our disposal. Effect sizes will be computed as follows, using the example of the primary outcome analysis. First, we will estimate an empty model as equation [4] above and obtain the unconditional sample estimates of the variances at the pupil σ_ε^2 , class σ_j^2 and school σ_k^2 levels. From the impact regression – in this example equation [1] above, we will obtain the adjusted sample estimate of the intervention effect $\hat{\beta}_1$. The effect size is then calculated as follows:

$$g = \frac{\hat{\beta}_1}{\sqrt{\sigma_k^2 + \sigma_j^2 + \sigma_\varepsilon^2}} \times \frac{N - 3}{N - 2.25} \times \sqrt{\frac{N - 2}{N}}$$

The second and third terms adjust for small sample bias and will have a trivial effect on the calculation in this case and will be ignored. Each effect size estimate reported will be accompanied by an uncertainty estimate in the form of a 95% frequentist confidence interval. This will be computed by taking the upper and lower limits of the confidence interval for $\hat{\beta}_1$ obtained from the regression output and dividing these limits by $\sqrt{\sigma_k^2 + \sigma_j^2 + \sigma_\varepsilon^2}$.

References

- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, 51(6), 1173–1182.
- Bell, R. M., & McCaffrey, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, 28(2), 169–182.
- Breaux, K. C., & Lichtenberger, E. O. (2016). *Essentials of KTEA-3 and WIAT-III assessment*. John Wiley & Sons.
- Carril, A. (2017). Dealing with misfits in random treatment assignment. *The Stata Journal*, 17(3), 652–667.
- 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. <https://doi.org/10.3389/fpsyg.2016.02056>
- 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. (2022). Statistical analysis guidance for EEF evaluations. <https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/evaluation-guidance-and-resources/evaluation-design>
- Gellen, S, Wicker, K, Ainsworth, S, Morris, S., & Lewin, C. (2024). *Using the DreamBox Reading Plus adaptive literacy intervention to improve reading attainment, a two-armed cluster randomised trial* [Study Protocol]. London, Education Endowment Foundation. https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/reading_plus_2024-25_trial_-_evaluation-protocol.pdf?v=1721601315
- Gerber, Alan, S., & Green, Donald, P. (2012). *Field experiments: Design, analysis, and interpretation*. W. W. Norton & Company.
- Imai, K., Tingley, D., & Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(1), 5–51. <https://doi.org/10.1111/j.1467-985X.2012.01032.x>
- Kaufman, A. S. & Kaufman, Nadeen, L. (2014). *Kaufman Test of Educational Achievement | Third Edition*. Pearson.

- <https://www.pearsonclinical.co.uk/store/ukassessments/en/kauffman/Kaufman-Test-of-Educational-Achievement-%7C-Third-Edition/p/P100009106.html>
- Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, 49(4), 241–253. <https://doi.org/10.3102/0013189X20912798>
- Lee, D. S. (2009). Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. *Review of Economic Studies*, Oxford University Press, 76(3), 1071-1102.
- Lewin, C, Morris, S., Ainsworth, S, Gellen, S, & Wicker, K. (2024). *Peer Assisted Learning Strategies UK (PALS-UK): A whole class reading approach—Evaluation Report*. London, Education Endowment Foundation.
- MacKinnon, J. G., Nielsen, M. Ø., & Webb, M. D. (2022). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics*. <https://doi.org/10.1016/j.jeconom.2022.04.001>
- McCarty & Ruttle. (2018). *Progress in Reading Assessment manual (Stage 2) (Second)*. Hodder Education.
- Moerbeek, M., & Schie, S. van. (2019). What are the statistical implications of treatment non-compliance in cluster randomized trials: A simulation study. *Statistics in Medicine*, 38(26), 5071–5084.
- O'Hare, L & et al. (2019). *Reciprocal Reading: Evaluation Report*. London, Education Endowment Foundation.
- https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Reciprocal_Reading.pdf?v=1695111397
- Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.
- Radach, R., & Kennedy, A. (2013). Eye movements in reading: Some theoretical context. *The Quarterly Journal of Experimental Psychology*, 66(3), 429–452.
- Spichtig, A. N., Gehsmann, K. M., Pascoe, J. P., & Ferrara, J. D. (2019). The Impact of Adaptive, Web-Based, Scaffolded Silent Reading Instruction on the Reading Achievement of Students in Grades 4 and 5. *The Elementary School Journal*, 119(3), 443–467.
- StataCorp LLC. (2023). *STATA CAUSAL INFERENCE AND TREATMENT EFFECTS ESTIMATION REFERENCE MANUAL*. Stata Press.
- Tauchmann, H. (2014). Lee (2009) treatment-effect bounds for non-random sample selection. *The Stata Journal*, 14(4), 884–894.

Tracey, L, Elliot, L., Fairhurst, C, Mandefield, L, Fountain, I, & Ellison, S. (2022). *Lexia Reading Core5®: Evaluation Report*. London, Education Endowment Foundation.

<https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Lexia-report-unconditional-effect-sizes.pdf?v=1695111274>

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2012).

Students with Learning Disabilities intervention report: Peer-Assisted Learning Strategies [WWC Intervention Report]. US Department of Education, IES. <http://whatworks.ed.gov>

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