PROJECT TITLE	Accelerated Reader
DEVELOPER (INSTITUTION)	Renaissance Learning
EVALUATOR (INSTITUTION)	RAND Europe and University of Cambridge
PRINCIPAL INVESTIGATOR(S)	Alex Sutherland
TRIAL (CHIEF) STATISTICIAN	Alex Sutherland
SAP AUTHOR(S)	Alex Sutherland and Sonia Ilie
TRIAL REGISTRATION NUMBER	ISRCTN17400885
EVALUATION PROTOCOL URL OR HYPERLINK	*https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/accelerated-reader-effectiveness-trial/*

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0	03/05/18	*Original*

Table of Contents

Introduction

Description of the intervention

Accelerated Reader is a whole-class reading management and monitoring programme that aims to foster the habit of independent reading among primary and early secondary pupils. The internet-based software initially screens pupils according to their reading levels (defined within the Accelerated Reader framework as the 'zone of proximal development' ZPD score), and suggests a ZPD score which helped guide pupils to appropriate books for their reading age and reading interest. Pupils take computerised quizzes on the books they have read and earn Accelerated Reader points; these points are related to difficulty and length of book read and also to the proportion of correct answers for each quiz and focus on the comprehension of the written text.

Schools receive three one-hour remote training sessions and six hours of training for staff implementing Accelerated Reader. These training sessions help schools ensure that their libraries are prepared for the intervention, show teachers how to use quizzes and how to analyse the data that they generate. Although requiring involvement of librarians or equivalent postholder within schools, Accelerated Reader does not have to be implemented by the whole school, and teachers can use it on their own in their classroom, provided that the library and IT systems are in place.

The intervention logic for the "Renaissance Learning: Accelerated Reader programme" (hereafter AR) involves: (i) the tailoring of assigned reading to pupils; (ii) the importance of reading within a pupil's zone of proximal development which includes reading at their current level and so-called 'stretch' books that are just above their current level (e.g. Vygotsky, 1978); and (iii) a constant feedback loop between reading and assessment, all designed to increase the amount of time children spend reading at school and thus ingrain the habit of reading.

The use of a standardised online platform to assist with initial screening and on-going testing means that there is a much lower risk of variation in how these work in practice (training or IT problems notwithstanding). But, evaluations of AR and discussions with the developers also suggest a more complex intervention. For example, AR also requires (iv) the initial selection of pupils and a range of book levels appropriate to them; (v) adequate resources in terms of books and IT infrastructure; (vi) protected and adequate amount of time for students to read on a daily basis (often on school premises) that often requires schools to change timetables (unless they already timetable dedicated reading time); and (vii) teachers to guide readers – especially the youngest and weakest – on book selection within their zone of proximal development. There is an element of (viii) staff training required, meaning that there is a need to understand how well the training has worked. Further, a previous EEF evaluation of AR (Gorard et al., 2015) illustrated a high degree of variability in how the additional reading was actually implemented across schools– and this is effectively the largest part of the intervention. Taken together, these elements suggest that schools' understanding of how to implement the programme, how effectively schools were able to schedule reading time, and how the intervention was 'sold' to pupils and parents could all affect treatment outcomes.

Study design

The trial is a two-arm, wait-list, cluster-randomised controlled trial, with schools as the unit of randomisation, and pupils as the unit of outcome analysis. The aim of the trial is to assess the extent to which Accelerated Reader leads to an improvement in reading ability (see outcome measures below) that is greater than what is observed in business-as-usual classrooms. The wait-list element relates to schools being randomised to receive AR in the 2016/17 school year, or the 2017/18 school year. Year 4 and 5 pupils in treatment schools will receive AR in the school year 2016/17. For control schools, it will be business as usual during the school year 2016/17. Year 3 and Year 4 pupils in control schools will eventually receive AR in 2017/18. This is akin to a cross-over design as control schools will receive the treatment, but those in the control year groups will not be eligible for AR (and compliance will be monitored using routine data collected by the developers).

Update: Intervention schools will receive the Intervention in the 2017/2018 year as well; please see the Protocol for details.

We will assess outcomes for Year 4 pupils after the third year of the trial usingKey Stage 2 (KS2) assessments, at the end of the third year of the trial, when the initial Year 4 students will be in Year 6.. Year 5 treatment effects will be assessed after two years using Key Stage 2 (KS2) assessments at the end of the second year of the trial. In effect, we will be conducting two trials at once – one for Year 4 pupils, the other for Year 5 pupils (Figure 2).

We will know whether treatment schools continue to use AR in the second year of the trial, and thus which pupils received extended exposure to AR and through the process evaluation we hope to capture what systems are in place in treatment schools that do not continue using AR. Whilst control schools will receive AR in the second year of the trial, the idea is that the intervention will only be available to Year 3 and Year 4 pupils in those schools. This is so that the comparison groups (Year 4 and 5 at the beginning of the trial, Year 5 and 6 in the second year of the trial) do not receive the intervention. By carefully monitoring the pupils accessing AR in control schools in 2016/17 we are able to assess the extent of contamination in the control school year groups (Figures 2a and 2b).

Figure 2a: Trial Design summary (original)

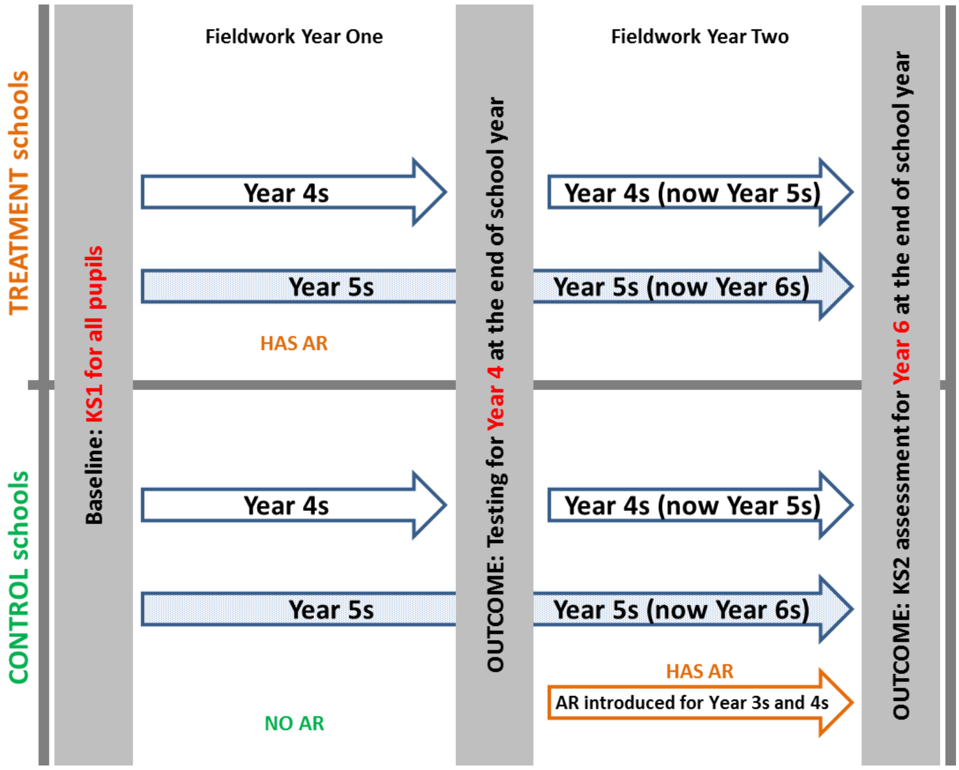
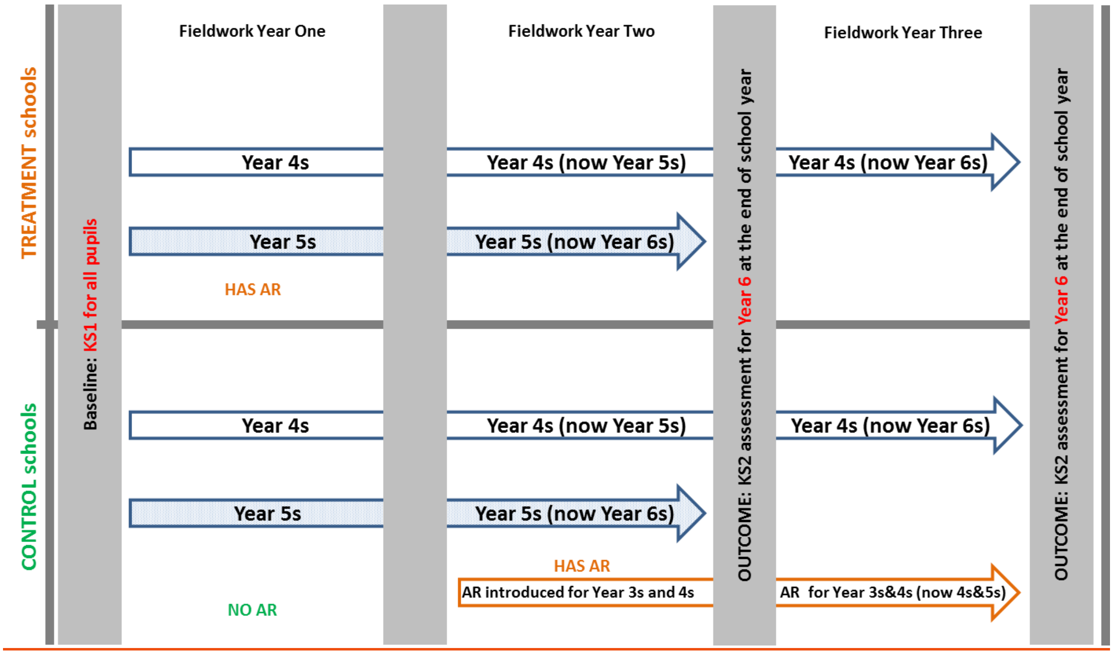


Figure 2b: Trial Design summary (post-extension)



School eligibility criteria

NFER[1] were responsible for recruiting schools in their role as delivery partner. As this is being funded as part of the Education Endowment Foundation's (EEF) North East Primary Literacy Campaign the EEF requirement for this trial was that approximately half of intervention schools were located in the North East while trying to ensure as high a proportion of FSM pupils as possible (but there was no specific target for FSM and this was not used during randomisation). Beyond that, any school that had used AR in the last two years was excluded.

Randomisation

A stratified cluster-randomised wait-list design was planned and undertaken, with schools as the unit of analysis. Allocation to treatment or control was conducted on the 5th of May 2016 by Dr Sutherland, with advice from Prof Anna Vignoles. Cluster randomisation was the preferred option (whereby the randomisation is done at the school level) to avoid the potential contamination effect of having intervention and control classes within the same school, given that some of the changes required to schools (such as the labelling of library books in accordance to AR level) could potentially spill over into the control condition. We ruled out one-shot simple randomisation because of the risk of imbalance on key factors such as geographical location. NFER recruited 181 schools for randomisation, from an initial target of 200 schools. These 181 schools were allocated to the intervention and control groups on a 1:1 ratio through stratified randomisation.

Strata were created by creating groups combinining **geographical region** (based on government regions) and a dichotomised measure of **school-level prior KS1 attainment**. KS1 prior attainment data was taken from the DfE website and used to mean-split the data, with the threshold being a KS1 average points score (KS1aps) of 15.2. Schools were categorised as 'low' for prior attainment if they had below 15.2 KS1aps (the sample mean was 15.098), and 'high' if they were equal to or greater than 15.2 KS1aps. Two schools did not

[1] National Foundation for Educational Research

have KS1aps data but were still included in the randomisation as a separate stratum. Regions were grouped together as follows:

- North East
- East + West Midlands ("Midlands")
- NW/Merseyside & South East

Within each region-KS1aps stratum, simple randomisation was conducted using a random number generator (in Stata). Each school within a stratum was assigned a random number, with schools then sorted by the random number and the first half in each strata being allocated to treatment. Random allocation resulted in 89 schools in the treatment group, and 92 schools in the control group. Table 1 below shows the allocations by strata:

Table 1: allocations by Strata

Region	KS1 high or low	Control	Treatment	Total
NE	Low	17	16	33
NE	High	15	15	30
Midlands	Low	12	11	23
Midlands	High	17	17	34
NW/M/SE	Low	16	15	31
NW/M/SE	High	14	14	28
	Missing KS1	1	1	2
Total		92	89	181

The evaluation team informed NFER and Renaissance Learning of allocations on the 5th of May 2016. As discussed below, trial analysis will incorporate stratum variables.[2]

Power calculations and sample size

The trial is intended to focus on analysis for all-pupils as the main effects analysis, but is powered for sub-group analyses of FSM-eligible pupils. We will also conduct analyses for non-FSM pupils, in keeping with proposed changes to EEF evaluation guidance.[3]

The aim was that the trial would consist of 200 schools, 100 in each treatment arm. This was the agreed upper limit placed on the trial in terms of both funding and the capacity of the developer to install and implement AR in time for a September 2016 start. Following EEF guidelines, we assumed that there are on average 33 pupils per class (66 per school) and an average Ever-FSM rate of 28%-29% as reported in DfE statistics (2015). With the assumed (average of) 33 pupils per class and 200 schools, the overall sample for Year 4 study is then 6,600 pupils (66 x 100) and another 6,600 pupils for Year 5, including approximately 2,000 Ever-FSM pupils in each Year. Based on the Intra-cluster correlation coefficient (ICC) analysis using 2014-15 NPD data completed by the EEF, we assume an ICC of 0.13. Finally, for comparability across trials, only pupil-level covariates will be used; specifically, based on KS1-KS2 correlation of r=0.73 (Treadaway, 2013), we assume level 1 variance explained of 0.73^2 = 0.53. We also assume an alpha of 5% and an intended 80% power to detect effects. Power and minimum detectable effect size (MDES) calculations were performed using PowerUp! (Dong and Maynard, 2013).

Table 2 sets out the MDES calculations for (i) the main analysis of all pupils and (ii) the analysis of FSM-eligible pupils in study schools and (iii) the non-FSM eligible pupils. For the planned

[2] Strata help to reduce differences between treatment and control groups so omitting them increases between group differences at baseline and a subsequent loss of power (see Kahan, B. C. and Morris, T. P. (2012) 'Reporting and analysis of trials using stratified randomisation in leading medical journals: review and reanalysis', *BMJ*, 345. http://www.bmj.com/content/bmj/345/bmj.e5840.full.pdf).

[3] These changes have not been confirmed as yet but we have been asked to include power calculations and update analysis plans to include this group.

analyses for all pupils – again based on EEF standards – the planned MDES was 0.15. For the analysis of FSM pupils the MDES was anticipated – based on standard EEF power calculation assumptions – to be 0.17. Finally for non-FSM pupils, assuming 2/3 of each class were non-FSM (22 pupils) and keeping all other parameters the same, the MDES was estimated to be 0.16 for the randomised sample (not tabled). If achieved, these effect sizes would compare favourably to the results from the previous EEF evaluation (Gorard et al., 2015: 5), which reported effect sizes of 0.38 and 0.24 for FSM and all pupils respectively.

Table 2: MDES calculation for each year group in the trial

		Protocol		Randomisation	
		OVERALL	**FSM**	**OVERALL**	**FSM**
MDES		0.15	0.17	0.16	0.18
Pre-test/ post-test correlations	level 1 (pupil)	0.73	0.73	0.73	0.73
	level 2 (class)	NA	NA	NA	NA
	level 3 (school)	0.00	0.00	0.00	0.00
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 tests?		Two	Two	Two	Two
Average cluster size		66	20	68.5[a]	22.8[b]
Number of schools	intervention	100	100	89	89
	control	100	100	92	92
	total	200	200	181	181
Number of pupils	intervention	6,600	2,000	6,274	2,032[c]
	control	6,600	2,000	6,124	2,101[c]
	total	13,200	4,000	12,398	4,133[d]

Table note: [a] Based on summing the average number of pupils per school for treatment and control arms (70.5 and 66.6 respectively) and dividing by two. [b] This is estimated by dividing the overall figure by three. [c] Based on multiplying the number of schools by the average number of pupils per school. [d] The sum of the two cells above.

It is worth remembering that the schools in the previous EEF trial had already opted to implement AR and as Gorard et al. (2015) point out, may be atypical in terms of their results. Similarly, those schools selected and randomised their own pupils and while Gorard et al. (2015) report the pupils were balanced, the randomisation was not observed by the evaluation team so there remains the chance of unintentional bias affecting results.

With the smaller sample of 181 schools allocated, the MDES is still below what the previous work (cited above) reports as achieved effect sizes for AR. We note that the calculation above uses estimated proportions of Ever-FSM-eligibile pupils, and that final analysis will include MDES calculation for the actual achieved Ever-6 FSM rate.[4]

[4] FSM-ever6 data are not yet available.

Actual Sample

NFER recruited 181 schools for randomisation. These schools were randomised (using procedure outlined above) to either the control or treatment groups. Following random allocation, there were 89 schools in the treatment group, and 92 schools in the control group as described above).

At the time of enrolment in the trial, the 89 intervention schools included 3,128 pupils in year 4, and 3,146 pupils in year 5. The 92 control schools included 3,076 pupils in year 4, and 3,048 pupils in year 5. Additionally, three pupils in the control schools were identified as year 2, and therefore not included in future analyses.

At the time of writing (June 2017) six of the 89 intervention schools had withdrawn from the trial, as had one control school. There were a range of reasons cited for withdrawal that will be fully reported following CONSORT standards, using weekly update information provided by NFER.

Outcome measures

Primary outcomes

The trial will be measuring two primary outcomes, one for each year group..

Year 4 pupils: the primary outcome measure will be the Key Stage 2 fine grade score for Reading, when the pupils are in Year 6 (end of the 2018-2019 school year). The measure will consist of the scaled fine grade scores, and will be collected from the National Pupil Database.

This represents a change to the original outcome measure intended for Year 4 pupils, as documented in the Trial Protocol update, and subsequent trial changes. This is due to data quality concerns in relation to the New Group Reading Test (NGRT) level 2A/2B, which was the initial outcome measure for Year 4 pupil.[5]

Year 5 pupils: the primary outcome measures will be Key Stage 2 fine grade score for Reading, when the pupils are in Year 6 (to examine the effect of AR over 2 years). These will be collected as scaled fine grade scores from the National Pupil Database (NPD).

Key Stage tests will be administered by schools in line with national guidance.

Secondary outcomes

No secondary outcomes will be analysed as part of the impact evaluation in this trial.

Analysis

Primary intention-to-treat (ITT) analysis

The analysis of primary outcomes (as detailed above) for both Year 4 and Year 5 pupils will be based on an intention-to-treat (ITT) framework.[6] This means that, once randomised, schools and participants will be analysed according to their allocation regardless of whether they received the intervention or not. The usual approach is to only include schools and participants with available outcome data.

[5] The final evaluation report will contain details of the problems with NGRT-II.

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

The ITT approach is particularly relevant for future policy-making stakeholders and practitioners, who may roll out or implement a particular programme, without much control as to how that programme is actually taken up in the system. Therefore, the ITT approach allows for estimating the effects of offering that particular programme.

The ITT approach is inherently conservative as it captures the averaged effect of *offering* the intervention, for those who do not comply and the effect for those who do comply. (This is not an issue for ITT analyses, but more for examinations of 'dosage'.) Another way of handling selective attrition – or selective exposure to treatment – that we will consider is the instrumental variables (IV) approach proposed by Angrist et al. (1996).

There is also the issue of student mobility once the trial has started (Vuchinich et al., 2012; Scochet, 2013).[7] Our approach would be to adhere to an intention-to-treat analysis in the event of students migrating between treatment and control schools. Pupils joining schools after the new school year had begun would be excluded from the evaluation but *would* be able to access AR.

Treatment and control arms will be compared in terms of the difference in means between groups at follow-up, conditional on baseline measures (KS1) and stratification variables.

It is important to note that cluster-randomised designs mean that both school and pupil level attrition may be possible post-randomisation. At the time of writing, school-level attrition from the trial has already occurred (see above). This has implications for analysis (see Schochet and Chiang, 2011),[8] although the schools that have dropped out from the trial post randomisation will still be included in the analysis as per the ITT framework as long as they have outcome data (this may be possible for Year 5 groups where the outcome measure is drawn from the National Pupil Database, see above, but will not be possible for Year 4 groups).

The unit of analysis is pupil level outcomes for those in schools with and without AR. There is an ongoing discussion about how 'best' to analyse results from RCTs that involve clustered data. Our approach is to 'analyse how you randomise', meaning that we will explicitly account for clustering via multilevel models (AKA 'random effects'). This approach assumes that the schools in the study are a random sample of all schools – which is often a source of contention – but one benefit of this approach is being able to explicitly partition variance and more flexibly handle complex variation within schools (Snijders and Bosker, 2012). Therefore, we will run our analyses using a multilevel approach, with pupils clustered in schools. The general equation for the multilevel model is given below:

$$y_{ij} = \propto + X_{ij}\beta + Z_i b_j + \delta AR_j + u_{ij} + u_j \qquad i = 1..N, j = 1..M, \qquad (1)$$

Where y_{ij} denotes the pupil level outcome; i and j denote pupil and school indexes respectively; X_{ij} is the $1 \times k$ vector of individual characteristics (here, prior attainment). AR_j is a dummy variable denoting treatment/control group at school level; β and δ are the $k \times 1$ and 1×1 vectors of regression coefficients; Z_j is a vector of school level characteristics, formed from the stratification variables mentioned previously. u_{ij} is the pupil-level error term; u_j is the school-level error term. In equation (1) α denotes the overall model intercept.

[7] **Schochet, P. Z. (2013). Student mobility, dosage, and principal stratification in school-based RCTs. *Journal of Educational and Behavioral Statistics*, *38*(4), 323-354.**

[8] While not widely known or reported, random effects models may yield biased estimates of ITT in cluster randomised trials under certain conditions when there is individual level noncompliance. Thus, it is critical to minimise individual level noncompliance and to include adequate covariates to reduce between-cluster variance. See, for example Jo et al. (2008).

The coefficient δ associated with the AR_j dummy will constitute the main result of the trial. "All pupils" represent the main analysis group (with the FSM-eligible and non-FSM eligible samples representing sub-groups), so to obtain the main result of the trial Equation (1) above will be estimated for all pupils.

The pupil characteristic that we will include is:

- Continuous variable for attainment at KS1, as the attainment baseline for both arms of the trial.

At the school level, we will include the two variables used to stratify schools: region and prior KS1 average points score. The former will be entered as n-1 dummy variables with the NE as the reference category. The KS1aps will be entered as a binary variable with 'low' as the reference category. As there were two schools without available KS1aps available at the point of randomistion, we will assess the impact on results of assigining both schools as 'low' and then as 'high' and finally omitting KS1aps from the model, as sensitivity analyses. If there is no substantive impact of the result for treatment versus control, we will specify how those two schools have been included, and report additional results as appendices.

These data will be obtained from two sources:

- Initially, and prior to randomisation, from schools: lists of pupils in the target year groups will be provided by all in-trial schools before schools are made aware of their allocation to control or treatment conditions;
- Secondly, from linkage to NPD, which will provide the first and last of the pupil characteristics above, as well as the primary outcome indicator (KS2 scores) for the second arm of the trial.

The actual relevance of statistical significance is debated and needs to be clearly defined. A p-value arising from a test of statistical significance represents the probability of observing a difference of the same size (or larger) in the population of interest in the event that the null hypothesis is correct (whatever the null represents). Here, the null hypothesis is that of 'no difference' between treatment and control groups. Study schools are sampled from a pre-defined population (detailed above), hence if the sampling is successful in terms of those schools being representative of the population they are drawn from, we will be able to make wider generalisations beyond the study sample.[9]

All analyses will be performed in Stata, versions 13.1 onwards.

Interim analyses

No analyses are planned for interim outcomes, but we will report results after both year two and year three as follows: results for Year 4 pupils (at that point Year 6) are collected and reported on at the end of the third year of the trial (2018/19 – reporting in 2019); results for Year 5 (at that point Year 6) pupils at the end of the second year (2017/18 – reporting in 2018).

Imbalance at baseline

A well-conducted randomisation will, in expectation, yield groups that are equivalent at baseline (Glennerster and Takavarasha, 2013). Because schools here are randomly allocated

[9] Even in the absence of a representative sample statistical testing would arguably be valid. With a well-executed RCT there should be baseline equivalence between treatment and control groups. These groups (and those sharing their characteristics) constitute the population of interest – hence testing for differences between the two groups (AKA 'in-sample testing') is a valid approach, as long as care is taken about the inferences made (see e.g. discussions in Hsu, 2008; Athey and Imbens, 2016).

to the control and intervention conditions, any imbalance at baseline will have occurred by chance. To check for, and monitor, imbalance at baseline, analyses will be conducted at the school and pupil level. At the school level, the analysis will look at the following variables, by means of cross-tabulations and histograms that assess the distribution of each characteristic within the control and treatment groups:

- Type of school (academy/non-academy)
- OFSTED rating
- Proportion of pupils eligible for FSM
- Proportion of pupils speaking English as an additional language (EAL)
- Number of pupils on roll
- School-level average KS1 scores (for sample pupils)

At the pupil level, the initial balance will be assessed for the following characteristics:

- Eligibility for FSM
- Gender
- KS1 attainment

Statistical significance tests will not be carried out to assess the balance, as their premise does not hold in randomised control trials[10] (i.e. given appropriate randomisation procedures were followed, any differences between control and treatment groups at baseline will be by definition due to chance, and classical statistical testing is therefore unnecessary). Instead, tables of the means (and standard deviation, where appropriate) for each characteristic will be presented, and the magnitude of any differences explored.[11]

An additional analysis will be carried out of aggregate pupil characteristics (gender, ever-6 eligibility for FSM, KS1 baseline attainment measure) at school level, to check for possible bias introduced due to attrition from the trial. This will take the form of an ANOVA (or equivalent regression analysis) of the analysed groups. This will only be relevant for the first arm of the trial whereby the primary outcome variable is not contained in the NPD.

Further analyses to be undertaken for the final report:

- Descriptive analysis of school characteristics, for both control and intervention groups, as at randomisation (following ITT framework, if and where outcome data is available, i.e. analysed schools will be those that have final outcome data)
- Descriptive analysis of pupil characteristics (gender, ever-6 FSM eligibility, baseline attainment measure, final attainment measures), in both control and intervention schools, as at randomisation (following ITT framework, if and where outcome data is available, i.e. analysed schools will be those that have final outcome data)
- Histograms of baseline and final attainment measures for control and intervention schools
- MDES calculation on the basis of achieved samples
- Analysis of the characteristics of the school and pupil sample in relation to the characteristics of all schools in England (available at aggregate level), to assess the representativeness of the study sample and potential for scale up.

[10] http://www.consort-statement.org/checklists/view/32-consort/510-baseline-data

[11] 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.

Missing data

Missing data can arise from item non-response or attrition of participants and / or schools. We will first determine the proportion of missing data in the trial, and then the shape of the missing data: that is, whether the data are missing completely at random (MCAR), missing at random (MAR), or whether there is systematic missingness; the former cases will affect the power of the trial; the latter case might suggest that bias will be an issue. Our use of administrative data for pupil baseline data and outcome data for Year 5 pupils, and the use of routine data collected via AR should reduce missingness arising from both item non-response and attrition. Should some data be missing, our analysis would always begin with an intention-to-treat (ITT) approach as set out above, meaning that we would look to analyse all schools and pupils as allocated at randomisation, regardless of compliance, and then move from there depending on the extent of missingness and where it occurs in the data.

We will explore attrition across trial arms as a basic step to assess bias (Higgins et al., 2011). We will provide cross-tabulations of the proportions of missing values on all baseline characteristics (as detailed in the previous section, at both pupil and school level), as well as on the primary outcome measures.

To assess whether there are systematic differences between those who drop out and those who do not – and thus whether these factors should be included in analysis – we will model missingess at follow-up using a multilevel logistic regression model as a function of baseline covariates, including treatment, and broader than those used for imbalance. For item non-response, the extent of missingness may in part determine the analytical approach.

For less than 5% missingness overall a complete-case analysis might suffice (i.e. assuming data are MCAR and that missingness is not assymetrical across groups), but our default will be to check results using approaches that account for missingness but that rely on the weaker MAR assumption. Our preference is to use Full-Information Maximum Likelihood (FIML) over multiple-imputation because FIML can be estimated in a single model and simulation studies show that it can reduce bias as well as MI (for a discussion of FIML vs MI see Allison, 2012). (For missingness on outcome variables only then standard statistical packages such as Stata use ML for estimating parameters so FIML would not be necessary (Allison, 2012).)[12]

Compliance-based analysis

The main framework of analysis for this trial is Intention-to-treat. However, we will also be able to explore the effect of the intervention on schools that were allocated to the intervention group and also followed allocation as intended (i.e. accepted the offered intervention). Following discussions with Renaissance Learning, we have agreed the following approach to capturing compliance in this trial, in order to assess to what extent those in the intervention arm have complied with allocation (i.e. actually implemented AR).

According to the developers: “The heart of the AR program is the extent to which teachers review and act upon reading practice data via AR”, that is, whether teachers intervene to change *what* and *how* pupils are reading, which the developers term *teacher responsiveness*.

[12] Allison, P. D. (2012) *Why Maximum Likelihood is Better Than Multiple Imputation*. Statistical Horizons. http://statisticalhorizons.com/ml-better-than-mi. And the more detailed discussion paper here: http://www.statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf.

Pupil fidelity is also regarded as important for AR's, defined by the developers as "To what extent are pupils in a school reaching a level of AR best practices implementation?".

The developers believe that both *teacher responsiveness* and *pupil fidelity* matter for AR's success. The rationale for looking at both is that just focusing on teacher responsiveness would mean that teachers who successfully implement AR in the first term (i.e. pupils hitting AR 'benchmarks') would be rated as 'not responsive' because they are not having to intervene. Hence, measuring pupil fidelity directly gives credit to teachers who have been more successful at implementing from the outset. It is important to note that the implementation of AR, particularly the emphasis placed on teacher responsiveness/pupil fidelity, blurs the distinction between 'compliance' and 'fidelity' because to be 'compliant' with AR, schools also need to be delivering with high fidelity to the intervention.

As part of our discussions with RL, using both measures would allow to decouple cases where teacher responsiveness and pupil fidelity varied, for example::

- "Schools with strong implementation throughout the year would receive high fidelity scores but low to moderate teacher responsiveness scores.
- Schools with poor implementation at the start but improvements throughout the year would receive low to moderate pupil fidelity scores but high teacher responsiveness scores.
- Schools with poor implementation throughout the year would receive low scores in both pupil fidelity and teacher responsiveness.
- Schools with mediocre fidelity and responsiveness would receive modest scores in both measures."[13]

To ensure consistency, we have agreed that RL will calculate measures based on term dates as follows in Table 3 (with half-terms excluded from calculations):

Table 3: Proposed School Terms: School Years 2016-17 and 2017-18

2016-17 School Year	2017-18 School Year
Term 1 (Autumn): 1st Sept - 31st December Term 2 (Spring): 1st January - 22nd April Term 3 (Summer): 23 April - 31st July	Term 1 (Autumn): 1st Sept - 31st December Term 2 (Spring): 1st Jan - 15th April Term 3 (Summer) 16th April - 31st July

RL have agreed to create measures of pupil fidelity and teacher responsiveness based on data available within AR (see Table 4). Pupil measurement is termly from Term 1, whereas teacher responsiveness is measured from Term 2 onwards (to give pupils time to us AR and teachers time to react).

Table 4: AR compliance measures

Compliance domain	**Calculated via…**
Pupil fidelity (measured termly from 1st term onwards)	a. Average percent correct (APC). Code each pupil as 1 where APC => 85%. Else, 0. Where APC is based on a pupil's average performance on AR quizzes across a given timeframe.

[13] Taken directly from communications with RL.

	b. Engaged reading time (ERT). Code each pupil as 1 where ERT => 15 minutes. Else, 0.[14] c. Zone of proximal development (ZPD). Code each pupil as 1 where average book difficulty level was within or above their ZPD. Else, 0.
Teacher responsiveness (measured from Term 2 onwards)	*Term 2 calculation*: Identify pupils from Term 1 coded as 0 in APC. Determine if their APC score in Term 2 was higher than the Term 1 score. If it was higher, code as 1, else 0. (The "1" indicates the teacher may have responded and intervened to improve performance or that the pupil changed without teacher intervention. The "0" indicates that there is no evidence of APC reaching the required threshold, but cannot tell us directly whether or not the teacher responded to the initial score.) Repeat for ERT and ZPD. Then, do the same for Term 3.

The pupil fidelity measures have previously been used to assess fidelity, and the developers have been clear that these measures are how Renaissance Learning assess compliance/implementation success. However, a concern raised by the EEF and discussed at length was that the pupil measures are 'treatment inherent' – i.e. are affected by intervention. Likewise, the indicator of teacher responsiveness is imperfect as can be capturing pupils' natural development (measured as an increase in APC) as well as teachers' actions. As such, this measure could also be treatment-inherent. The proposed approach may not circumvent these problems, but these were the agreed compliance measures section 2.7.1 of the original protocol.Other measures proposed by RAND and EEF were not acceptable to the developer as stand-alone measures. It remains a limitation of the research that some of the measures used for compliance may introduce upwards bias because they are treatment inherent.

Compliance score calculation steps

Table 5 sets out the calculation steps that RL will undertake as requested by RAND. Within each term (1, 2, and 3), RL will calculate the percentage of pupils meeting APC goals (those scored with 1 divided by the total number of pupils). This will be repeated for ERT and ZPD. Within each term (2 and 3), RL will calculate the percent of instances where there was evidence of teacher responsiveness for APC (those scoring a 1 in APC responsiveness divided by the number of pupils who had a 0 in APC pupil fidelity from the prior term). Then, do the same for ERT and ZPD. The total score will be calculated by summing the pupil fidelity and teacher responsiveness scores from each term. (Measures would be calculated using the data available for pupils for each term.)

[14] ERT is calculated using the following steps. Where ATOS is the reading difficulty of the book (taken from Renaissance Learning, 2015: 6).

$$points\ earned = \frac{10+ATOS\ level}{10} x \frac{words\ in\ book}{10{,}000}. \quad ERT = \frac{(AR\ points\ earned)\ x\ (minutes\ per\ point\ value)}{school\ days}$$

Table 5. Overview of Proposed Calculations

Term	Pupil Fidelity*	Teacher Responsiveness	Total Score
1	Term 1: (1a) Percent of pupils meeting APC target + (1b) Percent of pupils meeting ERT target + (1c) Percent of pupils meeting ZPD target	NA	1a + 1b + 1c
2	Same as above but limited to Term 2 (2a + 2b + 2c)	(2d) Percent of low fidelity APC pupils from Term 1 who increased APC in Term 2 compared to Term 1 + (2e) Percent of low fidelity ERT pupils from Term 1 who increased ERT in Term 2 compared to Term 1 + (2f) Percent of low fidelity ZPD pupils from Term 1 who increased ZPD in Term 2 compared to Term 1	Sum of (2a + 2b + 2c + 2d + 2e + 2f)
3	Same as above but limited to Term 3 (3a + 3b + 3c)	Same as above but focusing on Term 3 compared to 2 (3d + 3e + 3f)	Sum of (3a + 3b + 3c + 3d + 3e + 3f)
Total (Year)	Pupil Fidelity: Sum of the above (1a + 1b + 1c + 2a + 2b + 2c + 3a + 3b + 3c)	Teacher Responsiveness: Sum of the above (2d + 2e + 2f + 3d + 3e + 3f)	Total: Sum of all the above (1a + 1b + 1c + 2a + 2b + 2c + 2d + 2e + 2f + 3a + 3b + 3c + 3d + 3e + 3f)

We include a worked example of the calculations provided by RL in the Appendix to this SAP.

We will run an Instrumental Variables Model (IV) where treatment assignment is the instrumental variable and compliance with the treatment is the instrumented, endogenous variable. This will provide an indication of the effect of the intervention on ‘complier’ schools. Note that these analysis scale up the intervention effect by the share of those who complied.

Secondary outcome analyses

There are no secondary outcomes in this trial.

Subgroup analyses

The main outcome of the study will be the coefficient indicated by the δ in Equation (1) above when all pupils are included in the analysis. We will then treat FSM-eligible and non-FSM pupils as sub-groups and re-run analyses for each group separately after sub-setting the data.[15]

Based on our power calculations using the actual sample of 181 schools, the study should be able to detect a standardised mean difference (SMD) of 0.178 between FSM pupils in treatment and control groups. This is based on the above mentioned assumptions/parameters, specifically that there will be 29% of Ever-FSM pupils in the study population on average (i.e.

[15] There is a contradiction between the requirement to sub-set for FSM pupils and to run interactions for other sub-groups. ‘Evaluators should use appropriate statistical tests to check for interaction between the subgroups using the whole sample’ (EEF Evaluator Guidance, p.3). Whereas for FSM ‘The effect size on the primary outcome for FSM pupils should be calculated using a separate model’ (ibid. p.4). Guidance (e.g. Glennerster and Takavarasha, 2013) suggests using either strategy, but does not comment on whether mixing them is appropriate, but with interactions the caution is about the interpretation as this requires more care.

approximately 10 pupils in each class of 33 pupils) – a total of 2,000 Ever FSM pupils in each Year assuming a total population of 6,600 pupils per Year.

For the non-FSM pupils, the power calculations we have undertaken suggest a MDES of 0.16, within the same parameters as above.

The model for the sub-group (FSM-eligible and non-FSM pupils) analysis is provided below, in Equation (2).

$$y_{ij} = \propto + X_{ij}\beta + Z_j b_j + \delta AR_j + u_{ij} + u_j \qquad i = 1..N, j = 1..M, \text{(2)}$$

where y_{ij} denotes the pupil level outcome; i and j denote pupil and school indexes, respectively; X_{ij} is the $1 \times k$ vector of individual characteristics; ; Z_j is a vector of school level characteristics, formed from the stratification variables mentioned previously. AR_j is a dummy variable denoting treatment/control group; β, δ is the $k \times 1$ vector of regression coefficients; and u_{ij} is the error term.

Effect size calculation

With the multilevel models we will use the effect sizes for cluster-randomised trials given in the EEF evaluator guidance, an example, adapted from Hedges (2007) is given in Equation 1:

$$EQUATION\ 1{:}\ ES = \frac{(\bar{Y}_T - \bar{Y}_C)_{adjusted}}{\sigma^2}$$

Where $(\bar{\boldsymbol{Y}}_{\boldsymbol{T}} - \bar{\boldsymbol{Y}}_{\boldsymbol{C}})_{\boldsymbol{adjusted}}$ is the mean difference between intervention groups adjusted for baseline characteristics and σ^2 is the population standard deviation (variance), which we would request from the EEF and/or the Department for Education. The ES therefore represents the proportion of the population standard deviation attributable to the intervention (Hutchison and Styles, 2010). A 95% confidence interval for the ES, that takes into account the clustering of pupils in schools, will also be reported and will be calculated by dividing the upper and lower confidence interval bounds by the population standard deviation.

If the population variance is not available we would substitute the sample-based, pooled unconditional variance (s*) in the denominator, which is an estimate of the population variance. In multi-level models this variance will be the weighted average of the variance of treatment and control groups, as per EEF guidance (EEF, 2018: 4 footnoe 11). The formula is given in Equation 2:

$$EQUATION\ 2{:}\ s^* = \frac{(\bar{Y}_T - \bar{Y}_C)_{adjusted}}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}}$$

If there is reason to believe that AR has affected the spread of the distribution in the treatment group (dispersion),[16] the unadjusted variance of the control group would be substituted into the denominator.

[16] This would be assessed at looking at the distribution of outcomes pre/post for treatment and control groups and comparing statistical parameters of dispersion.

Appendix: example calculations for pupil fidelity and teacher compliance

Table A.1. Sample data (Term 1 & 2 only)

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
	Term 1						**Term 2**								
Pupil	T1 APC	(1a) T1 APC Fidelity	T1 ERT	(1b) T1 ERT Fidelity	T1 ZPD Avg (Range)	(1c) T1 ZPD Fidelity	T2 APC	(2a) T2 APC Fidelity	T2 ERT	(2b) T2 ERT Fidelity	T2 ZPD Avg (Range)	(2c) T2 ZPD Fidelity	(2d) T1-2 Resp APC	(2e) T1-2 Resp ERT	(2f) T1-2 Resp ZPD
1	70%	0	10	0	3.9 (2.4 – 4.5)	1	80%	0	18	0	4.1 (3.0 – 5.0)	1	1	1	-
2	80%	0	25	1	5.2 (4.5 – 6.5)	1	75%	0	22	1	5.0 (4.5 – 6.5)	1	0	-	-
3	82%	0	12	0	3.0 (2.4 – 4.5)	1	90%	1	12	0	3.0 (2.4 – 4.5)	1	1	0	-
4	90%	1	31	1	4.7 (2.4 – 4.5)	1	85%	1	30	1	5.2 (3.5 –5.5)	1	-	-	-
5	88%	1	20	1	2.2 (3.5 – 6.0)	0	90%	1	21	1	3.8 (3.5 – 6.0)	1	-	-	1
Summary Percent		.4		.6		.8		.6		.6		1.0	.67	.5	1.0

T1 = Term 1; T2 = Term 2

T1-2 Resp APC = Indication of Teacher Responsiveness. Did pupils with low fidelity in Term 1 improve their APC in Term 2? (1=Yes; 0=No; - is Not Applicable)

Summary indicates the proportion of 1 compared to all pupils with 1 or 0.

Sample narrative: Pupil 3's APC during Term 1 was 82%, just short of the target 85%+ so she is coded 0 for APC. Her ERT was 12 minutes, short of the 15-minute goal, so her ERT is coded 0. The average difficulty of books she read was 3.0, which was within her recommended ZPD of 2.4 – 4.5 so she is coded a 1 for ZPD. For teacher responsiveness, we consider those two areas she missed in Term 1: APC and ERT. We observe that her APC was up to 90% in Term 2, an improvement from Term 1, so her APC Responsiveness is coded 1. We observe that her ERT is 12 minutes in Term 2, which is the same as Term 1. Failing to improve that result results in coding ERT Responsiveness as 0. Since she met the ZPD goal in Term 1, we do not code ZPD Responsiveness in Term 2.

If the year ended after just two terms, and the school enrolled just five pupils, the summary would look like this:

Term	**Pupil Fidelity**	**Teacher Responsiveness**	**Total Score**
1	1a + 1b + 1c **.4 + .6 + .8 = 1.8**	*NA*	1a + 1b + 1c **.4 + .6 + .8 = 1.8**
2	2a + 2b + 2c **.6 + .6 + 1.0 = 2.2**	2d + 2e + 2f **.67 + .5 + 1.0 = 2.17**	2a + 2b + 2c + 2d + 2e + 2f **2.2 + 2.17 = 4.37**
3	3a + 3b + 3c *Not included due to space constraints*	3d + 3e + 3f *Not included due to space constraints*	3a + 3b + 3c + 3d + 3e + 3f *Not included due to space constraints*
Total	Pupil fidelity: **1.8 + 2.2 = 4.0**	Teacher responsiveness = **2.17**	Total Score: **1.8 + 4.37 = 6.17**