Evaluation Summary	
Age range	Primary school, Years 4 & 5
Number of pupils	Approximately 13,200[1]
Number of schools	200
Trial type	Effectiveness
Design	Wait-list cluster-randomised controlled trial
Primary Outcome	Reading Year 4: Key Stage 2 reading scores in 2018/19 school year Year 6: Key Stage 2 reading scores in 2017/18 school year

Summary of changes to Protocol

The main changes to the Protocol from the previously-published version, and detailed in what follows, consist of:

1. Extension of the trial, to now cover an additional school year. The end of the Intervention will now be June 2018 (end of the 2017-2018 school year). During the 2017-2018 school year, the Control schools will provide the Intervention to Years 3 and 4, as per initial protocol.
2. Change in outcome measure: for the initial Year 4 group only, the outcome measure has changed from the New Group Reading Test (NGRT 2A/2B) to the Key Stage 2 reading scores, measured at the end of the 2018/2019 school year, i.e. end of Year 6.

1. Background

1.1 Intervention

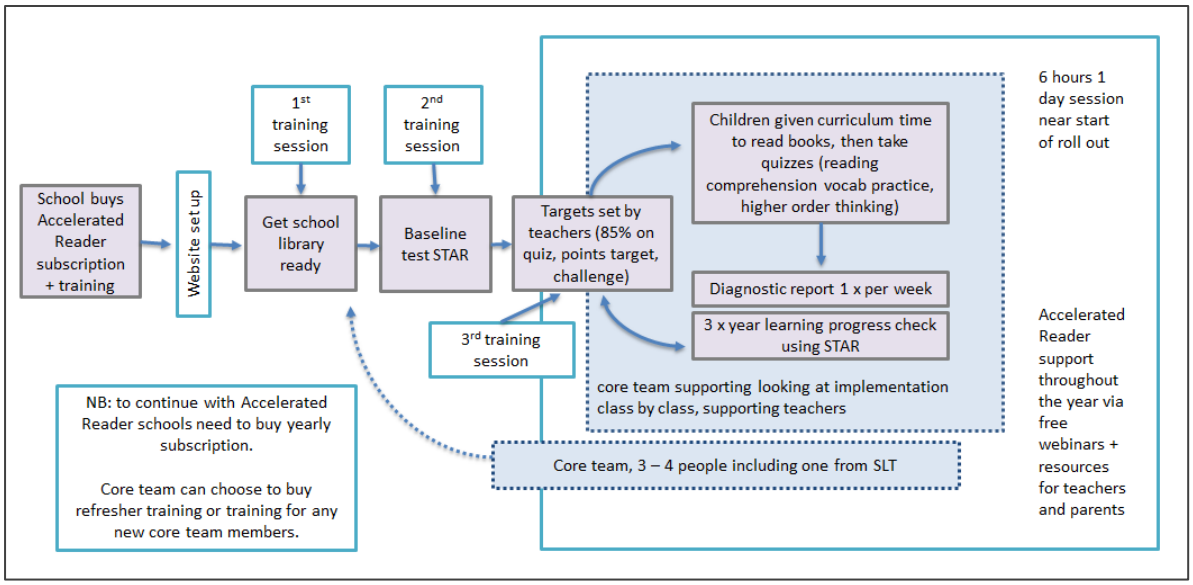
The intervention logic for the "Renaissance Learning: Accelerated Reader programme" (herein AR) involves: (i) the tailoring of assigned reading to pupils; (ii) the importance of reading within a pupil's zone of proximal development (e.g. Vygotsky, 1978); and (iii) a constant feedback loop between reading 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 also hint at 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); 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 this 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 – and this is effectively the largest part of the intervention. Taken together, these elements suggest that schools' understanding of how to

[1] This total assumes 200 schools and including Years 4 and 5 with an average of 33 pupils per year.

implement the programme, how effectively schools were able to schedule reading time, and how the intervention was 'sold' to pupils could all affect treatment outcomes.

Figure 1 below sets out the implementation logic model developed during pre-trial meetings between the Education Endowment Foundation (EEF), Renaissance Learning (RL), National Foundation for Educational Research (NFER) and the evaluators (RAND Europe and University of Cambridge).Figure 1: Accelerated Reader Logic Model



1.2 SIGNIFICANCE

The Renaissance Learning: Accelerated Reader programme is a widely used literacy programme which has been subject to some initial evaluation in the UK and elsewhere, including an aggregated trial funded by the EEF (See Gorard et al. 2015 and references cited therein, e.g. Topping 2014). This previous EEF evaluation was conducted by four secondary schools each doing their own within-school randomised control trial, the results of which were aggregated by an independent evaluator (Gorard et al. 2015). The study found that AR had a modest positive effect on the overall reading scores of the treatment group compared to the control group (effect size +0.24). When evaluated in terms of free school-meal (FSM)-eligible students (bearing in mind that they were not randomly assigned to groups), the effect was larger (+0.38), suggesting that AR may particularly improve the reading ability of Year 7 FSM-eligible pupils.[2] A recent summary of the evidence on AR for 'beginning reading' prepared for the US What Works Clearinghouse (Mathematica Policy Research, 2017) concluded that AR had 'mixed effects on comprehension and no discernible effects on reading fluency for beginning readers' (p.1). What was also notable was that some of the existing evaluations utilise the STAR reading test as an outcome, which is treatment inherent (see e.g. Appendix C1, p.16). Other AR evaluations carried outside of the UK have found positive, as well as, mixed effects of the programme (Ross et al. 2004; Ross et al. 2006; Nichols 2013; Huang, 2012; Shannon et al. 2015). For example, Ross et al. (2006) evaluated the reading attainment of 978 grade 3 to 6[3] students in a US urban school district. They found a large positive effect (+0.36) in the reading achievement of grade 3 pupils, while for grade 4 to 6 pupils the effect was positive but gradually smaller in older grades.

Here we will build on this evidence base with a randomised control trial in a larger number of UK schools. Additionally, this new RCT will focus on somewhat younger children, namely Years 4 and 5, in line with evidence of the benefits of very early intervention (Doyle et al. 2009). Consistent with the objectives of the EEF, and distinct from the previous evaluation, this new trial will be powered to detect an impact on disadvantaged children who are eligible for Free School Meals (FSM). This evaluation is also taking advantage of EEF work as part of the North East Literacy Campaign.[4]

2 Selection into treatment by FSM pupils is an issue faced by all EEF trials unless they are stratifying randomisation by FSM or are explicitly powered to detect effects for FSM pupils.
3 Equivalent to UK Years 4 to 7.
4 https://educationendowmentfoundation.org.uk/campaigns/north-east-literacy-campaign/

2. Methods

2.1 Research hypotheses

1. Hypothesis 1: Accelerated Reader will have a positive effect on the reading comprehension of all eligible pupils in the study year groups that are in randomly assigned intervention schools compared to all eligible pupils in control schools.
2. Hypothesis 2: Accelerated Reader will have a positive effect on the reading comprehension of pupils who are in the study year groups and are eligible for free school meals who are in randomly assigned intervention schools compared to FSM pupils in control schools.
3. Hypothesis 3: Accelerated Reader will have a positive effect on the reading comprehension of pupils who are in the study year groups and not eligible for free school meals who are in randomly assigned intervention schools compared to non-FSM pupils in control schools.

Note that the research hypotheses were updated at the request of the EEF to reflect a change in their approach to assessing FSM, non-FSM and all pupils.

2.2 Design

The trial will be 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 wait-list element relates to schools being randomised to receive AR for the first time 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. **Update:** Intervention schools will now receive AR in the school year 2017/18 as well.

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. We will assess outcomes for Year 4 pupils after the third year of the trial using Key Stage 2 assessments. Year 5 treatment effects will be assessed after two years using 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.

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 (as at 2016/17). By carefully monitoring the pupils accessing AR in control schools in 2016/17, 2017/18, and 2018/19 we are able to assess the extent of contamination in the control school year groups (see Figure 2a and 2b for the updated design).

Figure 2a: Initial Trial Design summary

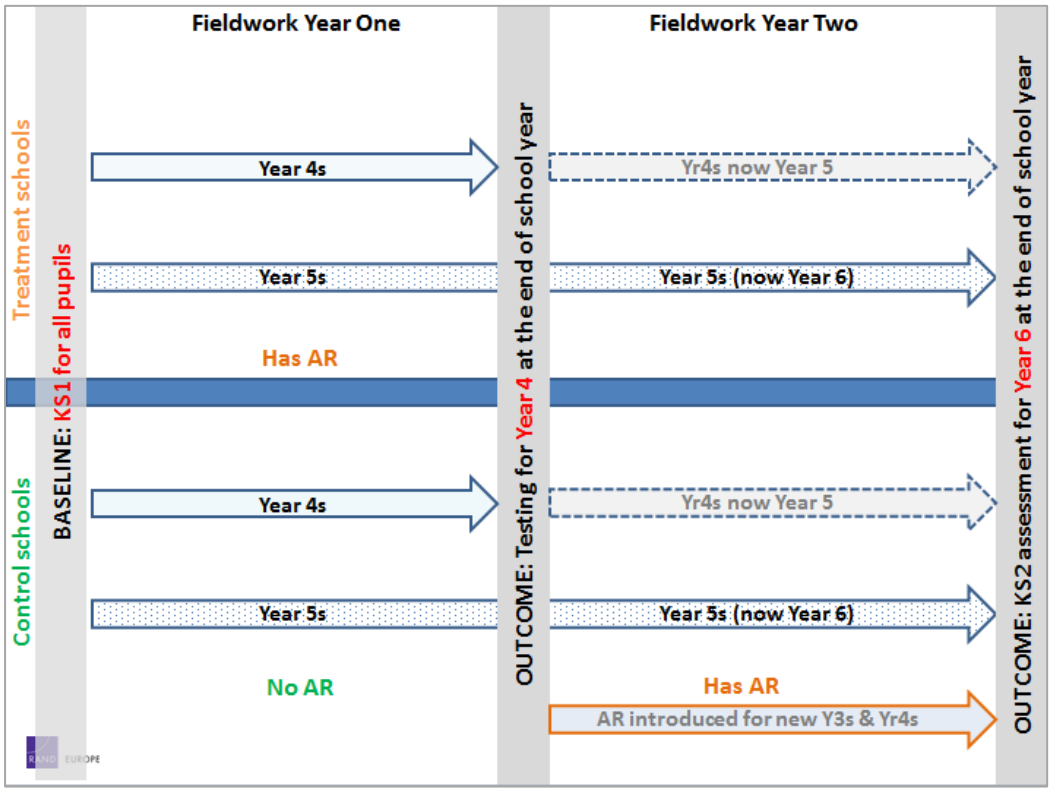
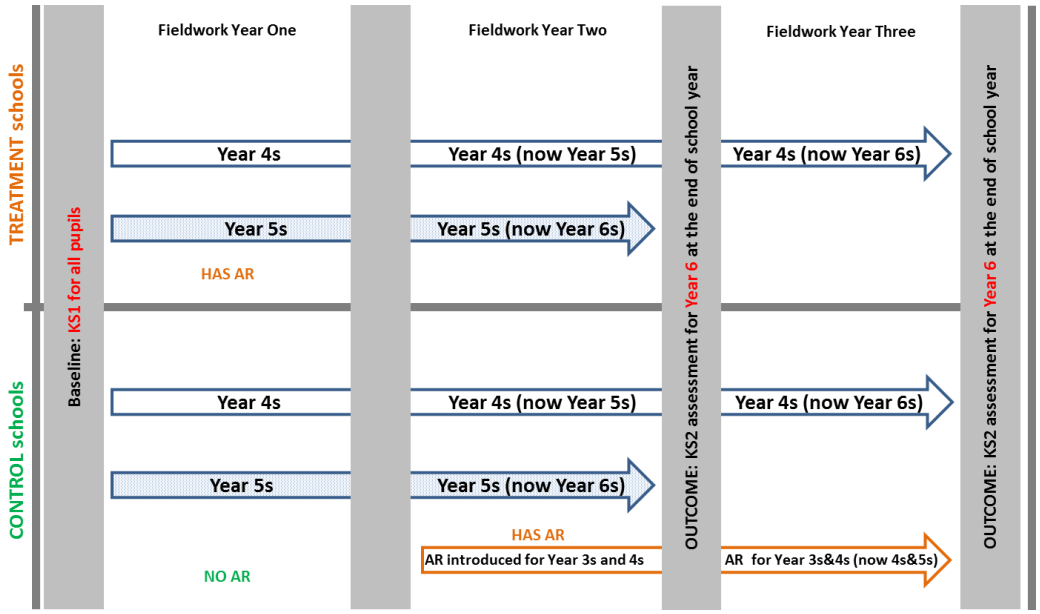


Figure 2b: Updated Trial Design summary



Protocol update: Following problems with the original outcome measure (described further below), it was agreed with the EEF that the intervention would continue for another year in the original treatment schools at the same time as allowing control schools to implement AR for the younger cohorts who were not part of the comparison group. To assess any contamination, the AR systems already in place (usage tracking) will be used. This means that pupils in treatment schools should receive six terms' worth of Accelerated Reader intervention.

The other change from the original trial plan was shifting the focal analysis from the Year 4 pupils to the Year 5 pupils (as they were when the trial began). Referring to Figure 2 above, Year 5 pupils entered Year 6 in the 2017/18 academic year, meaning that, as originally intended, KS2 test results could be used as the primary outcome measure. The effect of this change was to push the evaluation reporting timetable back a year, and lengthen the time the delivery and evaluation teams would be engaged with schools. Note that there was a period at the beginning of the 2017/18 academic year whilst this decision

was being made that some schools had discontinued their use of AR because they could not afford it. In total, 13 schools were affected by the discontinuation of AR (some up to a whole term), but they resumed once the decision was made. This represents under 10% of our initial total trial sample, and as a result the gap in usage is not expected to impact the trial overall. However, the consensus from delivery and evaluation teams was that gap should not materially affect the results of the evaluation.

2.3 RANDOMISATION

We will use a stratified randomised design and allocation will be conducted by members of the evaluation team. The exact set of stratification variables will vary depending on the availability of data, but we will reflect the sampling approach taken by NFER (discussed below) as well as factoring in any salient school characteristics such as the proportion of FSM students per year and prior KS1 results. We ruled out one-shot simple randomisation because of the risk of imbalance on key factors such as geographical location.[5] The strata will be created by creating groups based on sampling variables and other measures pertinent to pupil achievement. For example, through dichotomising variables based on median splits. Within each stratum, we will conduct simple randomisation using a random number generator (e.g. through that available in Excel, https://www.random.org/ or Stata). Each school within a stratum will be assigned a random number, with schools then sorted by the random number and the first half in each strata being allocated to treatment. As discussed below, trial analysis will incorporate stratum variables (and will assess the effects of including and excluding these measures on analysis). The evaluation team will inform NFER and Renaissance Learning of allocations. Schools will be informed of allocations by NFER in May 2016.

2.4 PARTICIPANTS

2.4.1 SCHOOLS

NFER will be responsible for recruiting schools in their role as delivery partner. As this is being funded as part of the EEF's 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 ensuring that the sample of schools had an average "Ever FSM" of 29%. Beyond that, any school that has used AR in the last two years will be excluded. Schools will be asked a series of questions about their existing facilities / reading lists. This is so that schools can be given a rough idea about how many books they might need to purchase, but also to assess the 'readiness' of the school for using AR as it is believed that the starting point of the school and the implementation of AR may be key factors in the effectiveness or AR.

The sample of schools was drawn from NFER's Register of Schools comprising of the following school types: primaries (including 'middle deemed primary', infant and junior, first- and middle, junior) maintained schools and academies. Special schools, PRUs and independent schools were excluded from the sample. The schools selected had pupils in all four of the following year groups: Year 3; Year 4; Year 5; Year 6. The schools selected were from the Local Authorities listed in the following table:

Table 1: geographical areas included in selection

LA Name	LA No	Notes
Birmingham City Council	330	
City of Newcastle upon Tyne	391	North East Area
Darlington Borough Council	841	North East Area
Derbyshire County Council	830	
Durham County Council	840	North East Area
Gateshead Council	390	North East Area
Halton Borough Council	876	
Hampshire County Council	850	
Hartlepool Borough Council	805	North East Area
Knowsley Metropolitan Borough	340	

5 Similarly, we also ruled out pairwise randomisation.

Liverpool City Council	341	
Middlesbrough Council	806	North East Area
North Tyneside Council	392	North East Area
Northumberland County Council	929	North East Area
Nottinghamshire County Council	891	
Redcar & Cleveland Borough Council	807	North East Area
Sefton Council	343	
South Tyneside MBC	393	North East Area
St Helens Council	342	
Stockton Borough Council	808	North East Area
Sunderland City Council	394	North East Area
West Sussex County Council	938	
Wirral Council	344	

The resulting sample comprised of 2,401 schools. A total of 206 schools were removed from the sample due to their being previous or current AR customers leaving a remaining sample of 2,195 schools available for invitation to take part in the evaluation.

STRATEGY FOR SCHOOL SELECTION AT THE RECRUITMENT STAGE

The target number of schools required to take part in the evaluation was 200. NFER estimated that approximately 10% of the sample of schools invited may volunteer to take part in the evaluation and so requested that the size of the sample for invitation contain around 2,000 schools.

Due to their being a finite number of (2,195) schools available for invitation it was necessary to contact all of the schools in the sample, rather than adopt a staggered approach whereby schools in the sample with a higher percentage of Ever FSM would be contacted before schools with a lower percentage of Ever FSM. In the event that an excess of schools are recruited over and above the 200 required, then it is proposed that the 200 schools that have the highest percentage of Ever FSM are selected and put forward for randomisation. The NFER and the evaluation team will assess the representativeness of the final sample compared to the population it was drawn from as this has implications for later generalisability and analysis.

2.4.2 PUPILS

All Year 4 and 5 pupils are eligible for inclusion in the trial, with the following exceptions. Schools using AR will use the AR STAR Reading test to determine pupils' zone of proximal development (measuring their reading range). STAR requires a sight vocabulary of 100 words, so students unable to take STAR would be excluded from the intervention. Furthermore, pupils unable to read independently at the lowest level of the zone of proximal development chart included in STAR would be excluded from the intervention.[6] For example, a Year 4 pupil able to read Year 1 books independently would be included in the trial. Based on Renaissance Learning's experience, it is believed that such pupils are typically unable to keep up with the requirements of AR, and so the intervention is not suitable for them. However, these pupils would still be able to use the AR system as to exclude them from the intervention would be unethical.

Pupils or their parents will be given the opportunity to opt out from the intervention.

2.5 OUTCOME MEASURES

The trial will be measuring primary outcomes for different year groups.

[6] AR includes three reading methods: read To, read With, and read Independent. Typically To and With are used with non-readers and developing-readers. A student who requires To or With can still use AR but may not be able to keep up with the requirements because a typical Y4 or 5 classroom teacher may not have adequate time to dedicate to only a few students.

Year 4 pupils: the primary outcome measure at the end of Year 4 (to examine the effect after 1 year) will be a measure of reading comprehension collected via the New Group Reading Test (NGRT) level 2A/2B (paper version). This test is available for Y2-Y4 pupils (ages 7-9), and lasts between 45-50 minutes. It provides standard overall scores that can be broken down into separate scale scores for sentence completion and context comprehension (GL-Assessment, 2015). It also delivers age equivalent scores and National Curriculum levels for reading (NGRT, 2015). For the trial analysis, the most fine-grained measure will be used, which in this case is the raw score on the NGRT Test 2A/2B.

Update to protocol: given the problems with the GL-Assessment NGRT-II data (detailed in the appendix), the primary outcome for Year 4 pupils was changed to be Key Stage 2 fine grade score for Reading (KS2_READSCORE) in the 2018/19 academic year.

Year 5 pupils: the primary outcome measures will be Key Stage 2 fine grade score for Reading (KS2_READSCORE), when the pupils are in Year 6, (2017/18 academic year) to examine the effect of AR after two years of intervention. These will be collected as scaled fine grade scores from the National Pupil Database (NPD).

All non-mandatory testing will be conducted / invigilated by NFER. Key Stage tests will be administered by schools in line with national guidance.

We intend to follow up the Year 4 pupils to assess the over-time impact of AR. The analysis for Year 4 pupils' Key Stage 2 data will be undertaken by RAND/Cambridge and reported as an addendum to the report.

To minimise assessor bias in outcome measures, allocation will only be shared with senior researchers from NFER who will not inform NFER field staff who will be overseeing test administration of school allocation. Assessor bias will also be minimised through the use of standardised and national testing of outcome measures.

2.5.1 SECONDARY OUTCOMES

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

2.5.2 BASELINE SCORES

We propose to use Key Stage 1 teacher assessment of the child's combined level in English (achievement in reading, writing and listening). These are part of a set of statutory assessments required for all pupils in maintained schools in England. Key Stage 1 tests have proved predictive of Key Stage 2 English tests and subsequently of overall GCSE achievement. Teachers administer the tasks and tests to the children and then determine the level the child is working at level 1, 2c, 2b, 2a, 3, or 4. Some children will be working towards level 1 and hence not achieve these levels. Although these tests are teacher assessed, the assessment and levels awarded will be determined prior to randomisation and indeed pre-trial, so would be suitable for use as baseline measures.

2.6 SAMPLE SIZE AND POWER CALCULATIONS

The aim is that the trial will 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. We assume that there are on average 33 pupils per class and average Ever FSM of 28%-29% as reported in DfE statistics (2015). With the assumed (average of) 33 pupils per year and 200 schools, the overall sample for Year 4 study is then 6,600 pupils 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 done 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 the PowerUp tool (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 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 – again 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 0.153. 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

Analysis	Initial calculation		
	All pupils	**FSM-ever pupils only**	**Non-FSM pupils**
Schools	200	200	200
Pupils per year	33	10	22
ICC	0.13	0.13	0.13
L2 Covariates	N/A	N/A	N/A
Variance explained by L1 covariates	0.53	0.53	0.53
MDES	**0.15**	**0.17**	**0.153**

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.

2.7 Analysis plan

The primary outcome for Year 5 pupils (after two years of AR) and for Year 4 pupils (after two years of AR and a one-year gap) will be reading attainment as measured by the Key Stage 2 exam results obtained via the NPD. 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. 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.

The unit of analysis here will be 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. One approach, 'analyse how you randomise', suggests that one should 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). Our approach will be to conduct sensitivity analyses to assess results against different model specifications[7]. These will be detailed in the Statistical Analysis Plan. The general equation for the multilevel model is given below as Eq.(1):

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

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 that include KS1 raw reading scores as a pre-test; Z_j is a vector of the stratification variables mentioned above; a_j is the school level (fixed) effect; b_j is the random effect; AR_j is a dummy variable denoting treatment/control group at the school level; β and δ are the $k \times 1$ and 1×1 vectors of regression coefficients; u_{ij} is the pupil-level error term; and u_j is the school-level error term. In equation (1) α denotes the overall model intercept.

The outcome analysis will be on an intention-to-treat (ITT) basis. Once randomised, schools and participants will be analysed according to their allocation regardless of whether they complied with the intervention or not. It is important to note that cluster-randomised designs mean that both school and pupil level attrition may be possible post-randomisation, with subsequent implications for analysis (see Schochet and Chiang, 2011).[8] 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.

7 For comparing fixed versus random effects approaches, the Hausman test assesses whether there are statistically significant differences between coefficients depending on model specification. However, model specification might also include the use of additional control variables.

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).

Problems of dropout/non-attendance may be an issue for this trial given that the intervention may require after-school attendance. (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). 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.

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]

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 below:

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

Where $(\bar{Y}_T - \bar{Y}_C)_{adjusted}$ is the mean difference between intervention groups adjusted for baseline characteristics and $\sqrt{\sigma_S^2 + \sigma_{error}^2}$ is an estimate of the population standard deviation (variance). The ES therefore represents the proportion of the population standard deviation attributable to the intervention (Hutchison and Styles, 2010). The exact effect size used will depend on whether there are equal or unequal sample sizes in trial arms.

2.7.1 SECONDARY ANALYSES

The study has been explicitly powered to detect an effect on FSM pupils. Based on the power calculations undertaken, the study should be able to detect a standardised mean difference (SMD) of **0.17** between FSM pupils in treatment and control schools. 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. approximately 10 pupils in each class of 33 pupils) – a total of 2,000 Ever FSM pupils each Year assuming a total population of 6,600 pupils per Year. The study will also report results for non-FSM pupils. The model used for the sub-group analyses of FSM and non-FSM pupils will be the same as Eq.(1) above, with the data sub-set for each group so that the approach is consistent.

The study will also assess the effects of treatment fidelity and 'dosage' on effectiveness for Year 4 pupils. Fidelity is how closely the actual intervention is to that intended. The basic argument that the greater fidelity is, the stronger treatment effects should be. Fidelity will be quantified via routine data collected by the Accelerated Reader programme. The primary measure(s) of fidelity for the trial will be: the percentage of treatment pupils offered AR that actually did take it up (e.g. 87%) and the percent of control pupils that were not supposed to access AR but did (e.g. 1.2%) (aka treatment adherence). The setup of AR also means that the 'dose' of AR will be driven both by pupil willingness and teacher effectiveness/willingness, neither of which are randomly assigned. This means that the intervention then becomes {AR + willingness/engagement}. Assessing a dose-response relationship in the absence of randomly assigned 'dosage' thus conflates random assignment treatment effects and selection effects unless it can be shown that dosage is exogenously determined. Further, AR measure three programme components for fidelity judged at the classroom level: Average Percent Correct on AR quizzes (should be 85% or higher), Engaged Reading Time (this metric is calculated within AR and should be around 30 minutes per day), and reading within the zone of proximal development.

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).

The cross-over of control schools into treatment also presents some risks and challenges. The first concern is that of contamination – here that Year 4 and Year 5 pupils in control schools will receive AR, violating the Stable Unit Treatment Value Assumption (SUTVA) thus breaking the randomisation (see Sinclair et al., 2012). To mitigate this, control schools would be offered AR *only* for their Year 3 and Year 4 pupils in the second year of the study. Use of AR by control pupils in what will be Year 5 and 6 pupils would be monitored, and schools would be reminded not to engage Year 5 and 6 pupils in AR. Renaissance Learning will provide information on use of AR in control schools so that non-compliance can be monitored.

Second, treatment schools will only have AR funded for one year, after which schools will then have to pay to continue with AR. As some schools in the treatment arm may opt to continue and others not, there may be differential 'doses' that could form the basis for exploratory analyses relating to dosage. Data for this analysis would be collated via the AR system and shared with the evaluation team. For those treatment schools not continuing with AR in the second year of the study, we would rely on the process evaluation survey to understand more about what practices were taking place (but this would likely be unable to inform the statistical analysis).

2.8 Missing data and attrition

Missing data can arise from item non-response or attrition of participants and / or schools. 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 approach as set out above, 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 baseline characteristics (to be specified in the statistical analysis plan) 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 would model missingess at follow-up as a function of baseline covariates, including treatment. 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), but our default would be to check results using approaches that account for missingness but that rely on the weaker MAR assumption. Our preference would be 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).[10]

2.9 Implementation and process evaluation methods

Following Oakley et al. (2006) and informed by both implementation science and realist approaches to evaluation, we will conduct a process evaluation to better understand the complexities of implementing AR. Implementation science emphasises thinking about 'dosage' or 'exposure', fidelity/adherence and implementation quality. The realist approach emphasises thinking about what works, for whom, and under what conditions. The purpose of the process evaluation will thus be to address the following questions:

- Was the intervention implemented with fidelity in the intervention schools?
- What factors and initial conditions appear to explain variation in fidelity of implementation?
- What appear to be the necessary conditions for success of the intervention?
- What were the barriers to delivery?

10 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).

Baseline contextual information will be examined for both control and intervention schools. This will include information about the current use of literacy-focused programmes and the current status of the school libraries.

The evaluation team will *observe the online training programme* for staff in the intervention schools. Each school receives three remote training sessions via WebEx/telephone and one onsite session. This will involve observing the three online sessions (by remote login), as well as observing the in-site session in three schools (to examine possible differences which might be due to different trainers in different locations). These observations will be accompanied by a review of the training materials and a review of the extensive online resources. Such observations will enable the team to better understand the aims of the training, its delivery, how it is initially received by the recipients, and the extent to which it matches up to the intervention logic model.

A *mid-intervention online staff survey* will be developed for the intervention schools. The survey will be administered to those who received the AR training (core team in each school), headteachers, year 4 teachers, TAs, librarians and literacy coordinators approximately half-way through the intervention period (i.e. early 2017). This survey will gather information on perceptions regarding the effectiveness of the training and resource materials, regarding the implementation of the intervention and any barriers or challenges faced by the staff in the implementation.

Following initial analyses of the mid-intervention survey, we plan to undertake workshops with staff in a sub-set of treatment group schools (i.e. in early 2017). These workshops will serve two main purposes. First, they will explore key points from the survey results to either validate or challenge survey findings through better understanding of underlying dynamics of the implementation of AR in practice. Second, the workshops will explore these issues in further depth than is possible through a survey, by encouraging workshop participants to elaborate on issues identified in the early stages of the process evaluation. This will include questions on teachers' experiences of the AR training sessions and the AR help resources, staff's experiences with the implementation of the intervention (and any barriers to the implementation of the intervention) and staff's perceptions of students' experience with AR.

The workshops will be located in areas with more than one (and ideally three to five) intervention schools in nearby commuting distance of one another, and we would invite two to three staff members from each school to attend. We plan to ask one school to host, ideally the most centrally-located school or the one most easily accessible by transit. We do not expect to be able to randomise the teacher selection for these workshops, and would work through each school to identify appropriate participants who were both available and willing to participate. Participants would be provided with refreshments and, where appropriate, certain travel costs could be covered by the project to ensure adequate participation.

A *follow-up online staff survey* will further gather information after the completion of the intervention at the end of the first year. The survey will similarly be administered to headteachers, year 4 teachers, TAs, librarians and literacy coordinators and head-teachers in the intervention schools. Both surveys will be piloted with online focus groups and/or expert reviews prior to administration.

Similar to the approach used during the intervention (described above), *post-intervention workshops* will be held following initial analyses of the follow-up survey. These workshops will provide an opportunity to gather more in-depth information regarding staff's perceptions regarding the overall implementation of the intervention (and barriers to implementation) and of its effectiveness and impact.

In parallel to this, *an online survey for headteachers* in control schools will be developed and distributed at the end of the first year. This survey will be conducted to better understand the kinds of literacy programmes that are in place in the control group schools, and how and to what degree these differ from (or are similar to) the AR programme in practice and will provide indicative data that may explain any differences (or similarities) in performance identified between treatment and control group schools.

Administrative implementation data from Renaissance Learning will be used to validate findings regarding the fidelity of implementation, as specified in Section 2.7.1 above. The Accelerated Reader system releases various reports which record data on: the number of books read by each student in a specified timeframe, the difficulty or book level, student participation rates as the percent of students who took at least one reading practice test within the last 30 days, engagement time which estimates the number of minutes per day that a student was actively engaged in reading practice, and data on student comprehension represented by their attained test scores. The feasibility to obtain this

information in a single dataset with school level indicators, and to acquire additional data relevant to measure implementation fidelity, will be further discussed with RL. RL research suggests that indicators of Engaged Reading Time (which takes into account the number of points a student earns relative to a given timeframe), high percent correct on quizzes, and reading in the zone of proximal development correlate with the strongest achievement gains.

In the second year of the trial, the routine administrative data collected by AR will form the basis of understanding whether there has been any 'contamination' (i.e. whether the Year 5 and 6 pupils in control schools have accessed AR). The teacher/headteacher surveys will also be re-administered by NFER in both treatment and control schools (following suitable edits) to understand whether and how practice has changed in treatment/control schools. One interesting question is whether treatment schools continue to use AR – something that can be tracked using routine data. Similarly, in those schools that do not continue with AR the question will be what else they are doing, which is something we aim to capture in the surveys.

2.10 Reporting

The findings from this study will be reported in two separate outputs: (i) the main study report containing details of the study design, implementation, process data and results for Year 4 pupils. This will be reported in 2017. (ii) the Year 5 data using KS2 data and reporting on updated process evaluation data (surveys), this will report in 2018/2019 depending on the timing of KS2 being released by the DfE.

2.11 Costs

We will estimate average marginal costs per pupil per year for schools for the intervention. We will do this in two stages: an initial discussion with the developer and then asking headteachers to populate a template, outlining key cost categories in both start-up and operation of the AR intervention. Average marginal costs per pupil per year over a three year period will be estimated in line with the convention followed by other Education Endowment Foundation projects. Separate estimates of prerequisite costs and of additional staff time will also be provided. These estimates together will provide a general indication of the costs involved with implementing AR, though we recognise that costs may differ substantially between schools and in particular pre requisite costs will vary substantially depending on the existing library resources and books available.

Specific cost categories on which we will collect data include:

1. Direct financial costs, including direct costs paid to AR if the intervention were to be offered without subsidy, salary costs for additional staff required, direct costs for books and other materials etc.
2. Direct time costs, including staff time and volunteer time. This will be reported separately. We will also report average supply cover time (rather than cost) required for the intervention.
3. Prerequisite costs, such as the costs of establishing an appropriately stocked library.

We will collect additional data from the delivery organisation to validate the above estimates and to ascertain whether any changes to costs are anticipated in the future if the intervention were to be made more widely available.

4. Ethics and registration

International Standard Randomised Controlled Trial Number (ISRCTN): ISRCTN17400885.

Following the EEF Consent and the Data Protection Act document, and as a result of discussions between the EEF and Department for Education, the trial will require school-level consent and parental opt-out. With participants' parents or legal guardians being the decision-makers for individual pupils. The main reason for this is that the intervention will be delivered within the school day when schools act *in loco parentis*, in the place of a parent, and the intervention does essentially not differ from the standard practice in schools.

In terms of fair processing of personal data, the project will fulfil the Condition 1 of processing personal data in Schedule 2 of the DPA as the data subjects will give their implicit consent in form of an opt-out letter to parents at the beginning of the trial. The ethics and registration processes are in accordance with the ethics policies adopted by RAND Europe. The study was reviewed by the RAND Europe ethics advisory board and approved by the University of Cambridge, Faculty of Education ethics review process.

RAND Europe adopts good industry practices regarding the protection of personal data as part of its obligations as a Data Controller under the Data Protection Act 1998 and takes appropriate technical and organisational measures conformant with ISO 27001 to protect personal data. Individuals targeted by the study have the right to oppose, have access to, rectify, or remove personal or sensitive personal data held by RAND Europe and University of Cambridge.

Data sharing between RAND and Cambridge will be governed by a data sharing agreement.

5. Personnel

There are several organisations involved in this evaluation. Table 3 below sets out the organisational roles and responsibilities.

Table 3: Organisational roles and responsibilities for AR evaluation

Organisation	Role	Responsibilities
RAND Europe & University of Cambridge	Evaluators	Overall responsibility for the outcome and process evaluations, trial design, analysis, reporting and quality assurance of the study.
NFER	Recruitment & testing	• Recruiting a random sample of schools within the constraints placed on the trial. • Regular reporting on recruitment progress. • Documenting school and pupil recruitment / flow according to CONSORT reporting requirements (Campbell et al., 2010). • Collating a list of schools and pupils included in the trial that allow for later secure data linkage. • Administering assessor blinded testing for all Year 4 pupils in treatment and control schools. • Ensuring that reasonable steps have been taken to reduce missing data for Year 4 pupils' outcome data. (e.g. through scheduling follow-up visits to schools to re-administer Year 4 outcome test for those missing on the day of testing). • Scoring and collating test data for Year 4 pupils in treatment and control schools. • Returning pupil-level raw and cleaned test data, along with the code used for data cleaning, for all Year 4 pupils in treatment and control schools once these have been completed. • Creating and maintaining a key contact list for all schools during the study and sharing this with the evaluator team. • Facilitating access to schools for evaluator team fieldwork. • Reporting on implementation of AR. • Distributing links to online surveys to headteachers and Year 4 & 5 classroom teachers.
Renaissance Learning	Intervention provider	To install the AR system in all schools as per the terms of the sign-up information sheet. To provide pupil-level data on any AR tests undertaken, and key indicators of implementation and compliance with the intervention.
Education Endowment Foundation	Funder	• Funding RAND-Cambridge, NFER and (via NFER) Renaissance Learning. • Facilitating set-up phase. • Reviewing reports. • Signing off on evaluation.

4.1 TEAM

Project Leader: Dr. Alex Sutherland (RAND) | *Project Managers*: Dr. Julie Bélanger & Dr. Megan Sim (RAND)

Core fieldwork and analysis team: Dr. Sonie Ilie (Cambridge) | Dr. Chris Giacomantonio (RAND, qualitative methods) | Miriam Broeks (RAND) | Martin Stepanek (RAND).

Project advisors: Professor Anna Vignoles (Cambridge, education research) | Jon Schweig (RAND, RCTs)

5. Risks

Risk	Assessment	Mitigation strategy
Recruitment failure	Likelihood: Moderate Impact: High	NFER to recruit schools and be the main contact for schools. NFER to allow sufficient window for recruitment. Clear information about benefits of participation. Use of wait-list design as incentives for control schools.
Attrition	Likelihood: Moderate Impact: Moderate to high	Clear information about expectations and requirements provided to participating schools. Intention to Treat (ITT) analysis to be used. Attrition to be monitored and reported according to CONSORT guidelines (Campbell et al., 2010).
Missing data	Likelihood: Moderate Impact: Moderate	Use of administrative data for Year 5 pupils. Follow ups with pupils missing on the day of assessment for Year 4 pupils.
Student mobility	Likelihood: Moderate Impact: Low	Pupils who are included in the study at the start of the school year and who move between study schools will be retained and analysed according to their original allocation to treatment / control. Year 4 pupils who migrate to non-study schools will be excluded from the Year 1 analysis as these pupils will be tested with external tests. In the event that mobility to non-study schools exceeds 10% on average across all schools, then the evaluators will discuss with the EEF the possibility of additional funding to collect this information. All pupils will be tracked using NPD data so all Year 5 pupils will be retained in the study and analysed according to original allocation.
Low implementation fidelity	Likelihood: Low to moderate Impact: Moderate	Process evaluation to monitor and document fidelity of implementation
Cross-contamination	Likelihood: Low Impact: High	School randomisation to be used. Schools with experience using AR in the past two years excluded from sample. Cross-over period to exclude pupils participating in the trial & this to be monitored via AR meta-data.
Evaluation team members absence or turn-over	Likelihood: Moderate Impact: Low	All RAND staff have a three month notice period to allow sufficient time for handover. The team can be supplemented by researchers with experience in evaluation from the larger RAND Europe pool.
Low participation	Likelihood: Moderate Impact: Moderate	School locations for workshops to be chosen based on convenience for participants to attend.

rates for workshops		Refreshments to be served during workshops.
Low response rates for surveys	Likelihood: Moderate Impact: Moderate	Surveys to be kept to a maximum of 10-15 minutes long. Respondents given the opportunity to complete survey online on multiple occasions if required. Sufficient data collection window given with real-time monitoring of response rates to allow for reminders to be targeted.
Lack of coordination with larger teams (RAND, Cambridge, NFER, RL, EEF)	Likelihood: Moderate Impact: Moderate	Teams to attend initial meetings and agree on roles and responsibilities at the outset. Regular updates to be provided to the lead evaluators. Regular contact between senior team from each organisation.

6. Data protection statement

RAND Europe adopts good industry practices regarding the protection of personal data as part of its obligations as a Data Controller under the Data Protection Act 1998 and takes appropriate technical and organisational measures conformant with ISO 27001 to protect personal data. Respondents to this survey have the right to oppose, have access to, rectify, or remove personal or sensitive personal data held by RAND Europe.

7. Timeline

Date	Activity
Jan 2016	Evaluation kick-off meetings
Feb-March 2016	Recruiting schools
April-May 2016	Randomisation
May-July 2016	Training provided for intervention schools
Sept 2016	Intervention begins in YR4 and YR5 in intervention schools
Jan 2017	Mid-intervention staff survey in intervention schools
Feb 2017	Mid-intervention workshops in intervention schools
June 2017	Outcome tests for YR4 pupils
June 2017	Follow-up survey with intervention schools
June 2017	Phone interviews with head teachers in control schools
July 2017	Workshops with intervention schools
Apr-July 2017	Training provided to control schools
Sept 2017	Intervention begins in YR3 & YR4 of control schools Intervention continues in (now) YR5 & YR6 of intervention schools
Jan 2018	Follow up survey with control schools
June 2018	Final year survey with intervention and control schools
Nov 2018	First report
Dec 2019	Second report (including younger cohort (YR4) KS2 outcomes)

8. Conflict of interest statement

None of the evaluation team has any conflicts of interest and all members of the study team have approved this protocol prior to publication.

References

Athey, S. and Imbens, G. W. (2016) 'The Econometrics of Randomized Experiments', chapter 5 in the *Abdul Latif Jameel Poverty Action Lab Handbook of Field Experiments* (eds Banerjee and Duflo) https://www.povertyactionlab.org/handbook-field-experiments. Cambridge, MA: J-PAL. (Last accessed 8th June 2016)

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.

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996) 'Identification of Causal Effects Using Instrumental Variables', *Journal of the American Statistical Association*, 91(434): 444-455.

Campbell MK, Piaggio G, Elbourne DR, Altman DG; for the CONSORT Group. (2012) 'Consort 2010 statement: extension to cluster randomised trials', *BMJ*, 4(345):e5661. PMID: 22951546

Department for Education statistics [DfE statistics] (2015). National Statistics: Schools, pupils and their characteristics: January 2015. Available at https://www.gov.uk/government/statistics/schools-pupils-and-their-characteristics-january-2015

Dong, N. and Maynard, R. A. (2013). 'PowerUp!: A tool for calculating minimum detectable effect sizes and sample size requirements for experimental and quasi-experimental designs', *Journal of Research on Educational Effectiveness*, 6(1), 24-67.

Doyle, Orla, Colm P. Harmon, James J. Heckman, and Richard E. Tremblay (2009) 'Investing in early human development: timing and economic efficiency', *Economics & Human Biology*, 7(1): 1-6.

Education Datalab (2015) *Seven things you might not know about our schools*. London: Education Datalab. http://www.educationdatalab.org.uk/getattachment/Blog/March-2015/Seven-things-you-might-not-know-about-our-schools/EduDataLab-7things.pdf.aspx

GL-Assessment (2015). New Group Reading Test – Sentence completion and passage comprehension in a single test. Available at http://www.gl-assessment.co.uk/sites/gl/files/images/Files/NGRTsub-scalepaper.pdf

Gorard, S., Siddiqui, N., Huat See, B. (2015) Accelerated Reader. Evaluation report and executive summary, February 2015. Available from https://educationendowmentfoundation.org.uk/uploads/pdf/Accelerated_Reader_(Final).pdf, last accessed 16th October 2015.

Hedges, L. V. (2007) 'Effect sizes in cluster-randomized designs', *Journal of Educational and Behavioral Statistics*, *32*(4): 341-370.

Higgins JP, Altman DG, Gotzsche P. C., Juni P., Moher D., Oxman A. D., Savovic J., Schulz KF, Weeks L, Sterne JA. (2011) 'Cochrane Bias Methods Group; Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials', *BMJ*, 343:d5928.

Hsu, L. M. (2008) 'Random Assignment Procedures', in Nezu, A. M. and Nezu, C. M. (Eds.) *Evidence-Based Outcome Research: A Practical Guide to Conducting Randomized Controlled Trials for Psychosocial Research*. Oxford: Oxford University Press.

Huang, S. (2012) 'A mixed method study of the effectiveness of the Accelerated Reader program on middle school students' reading achievement and motivation', *Reading Horizons*, *51*(3): 229-230.

Jo, B., Asparouhov, T., and Muthén, B. O. (2008) 'Intention-to-treat analysis in cluster randomized trials with noncompliance', *Statistics in medicine*, 27(27): 5565.

K., King, G. and Nall, C. (2009) 'The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation', *Statistical Science*, 24(1): 29-53. This approach is advocated but not commonly used in social science.

NGRT (2015) New Group Reading Test. Available at http://www.gl-assessment.co.uk/products/new-group-reading-test

Nichols, J. S. (2013) *Accelerated reader and its effect on fifth-grade students' reading comprehension* (Doctoral dissertation, Liberty University).

Oakley, A., Strange, V., Bonell, C., Allen, E., and Stephenson, J. (2006) 'Process evaluation in randomised controlled trials of complex interventions', *BMJ (Clinical research ed.)*, *332*(7538): 413-416.

Ross, S. M., Nunnery, J., & Goldfeder, E. (2004) *A randomized experiment on the effects of Accelerated Reader/Reading Renaissance in an urban school district: Final evaluation report*. Memphis, TN: Center for Research in Educational Policy, The University of Memphis.

Nunnery, J. A., Ross, S. M., & McDonald, A. (2006) 'A randomized experimental evaluation of the impact of Accelerated Reader/Reading Renaissance implementation on reading achievement in grades 3 to 6', *Journal of Education for Students Placed at Risk*, *11*(1): 1-18.

Schochet, P. Z., & Chiang, H. S. (2011) 'Estimation and identification of the complier average causal effect parameter in education RCTs', *Journal of Educational and Behavioral Statistics*, *36*(3): 307-345.

Shannon, L. C., Styers, M. K., Wilkerson, S. B., & Peery, E. (2015) 'Computer-assisted learning in elementary reading: A randomized control trial', *Computers in the Schools*, 32(1): 20–34.

Sinclair, B., McConnell, M. and Green, D. P. (2012) 'Detecting Spillover Effects: Design and Analysis of Multilevel Experiments', *American Journal of Political Science*, 56(4): 1055–1069.

Snijders, T. A. B. and Bosker, R. J. (2012) *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modelling* (2nd edition). London: Sage Publishers.

Topping, K. (2014) *What kids are reading: The book reading habits of students in British Schools 2014*: *An independent study*. Renaissance Learning Inc.: United Kingdom.

Treadaway, Dr M. (2013). *An analysis of Key Stage 2 reliability and validity*. FFT Research Paper No.2, FFT Education Ltd. Available from: http://www.fft.org.uk/FFT/media/fft/News/KS2_Reliability_-_Validity_MT_FINAL.pdf

Vuchinich, S., Flay, B. R., Aber, L., & Bickman, L. (2012). 'Person mobility in the design and analysis of cluster-randomized cohort prevention trials', *Prevention Science*, *13*(3), 300-313.

Vygotsky, L. (1978) 'Interaction between learning and development'. From *Mind and Society*. Cambridge, MA: Harvard University Press. Reprinted in Gauvain, M. and Cole, M. (eds) (1997) *Readings on the Development of Children* (2nd edition). New York: W. H. Freeman and Company.

WWC (2008) WWC Intervention Report: Accelerated Reader. Available from http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/wwc_accelreader_101408.pdf, last accessed 16th October 2015.

Appendix: problems with original outcome testing and changes to trial timetable

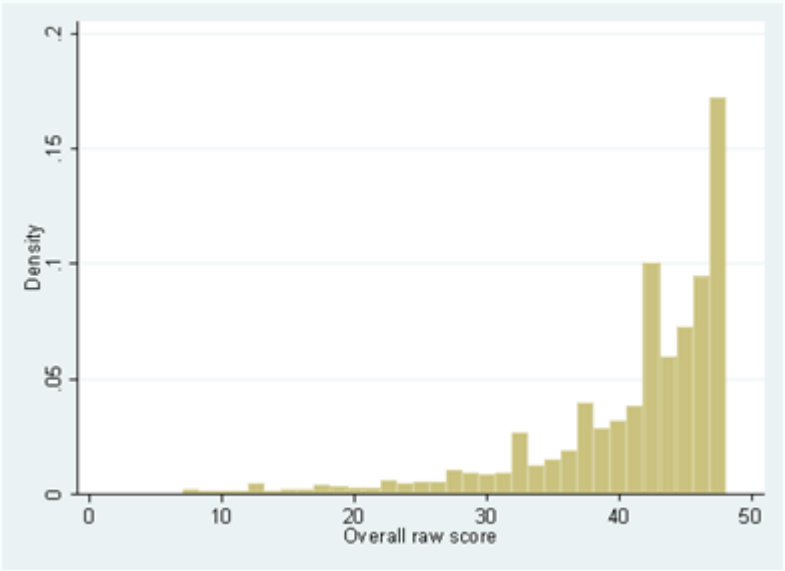
The original project plan was, as per Figure 2 above, that at the end of the 2016/17 academic year all Year 4 pupils would be tested using an external test, the New Group Reading Test II (owned by GL-Assessment). This was so that there would be results from the study more quickly, and because of the wait-list design the concern was that there would be contamination between treatment and control years.[11] The test was administered by independent invigilators from NfER and there were no reported problems with test administration.

However, when the test data from the NGRT-II were received, the distribution of the outcome data was strongly suggestive of a problem with the test. To understand the issues, Figure 3 shows the distribution of the test results for approximately 3,500 Year 4 pupils. The test was scored between 0-48. Ordinarily, one would expect to find a measure of educational ability or aptitude to be approximately normally distributed; with the majority of observations clustered around a mid-point and following a bell-curve. We observe that the majority of observations are distributed to the right, meaning that the majority of pupils scored highly on this test (around 30% of pupils scored 47/48 or 48/48). This fact gives rise to several possible explanations:

- The wrong test could have been being used.
- There were scoring errors / tests were incorrectly marked.
- There are so-called 'ceiling effects', wherein the test was too easy.

Having checked the scoring ourselves, we found this was correctly done. Similarly, having checked the physical tests themselves by means of the actual tests held by NFER, these were the correct ones – at least the test that was supposed to have been used. We also checked by looking at the commercial website for GL-A and followed their guidance on how to choose the right test for this age-group. This left us to conclude that the test was too easy for the pupils and thus could not differentiate between them.

Figure 3: NGRT-II test data for approximately 3,500 Year 4 pupils



The problem for the evaluation is that these data are not suitable for assessing the effectiveness of AR on pupil outcomes. Standard statistical testing relies on assumptions about distributions of differences

[11] The combination of being able to monitor use of AR very closely and implementing AR use in a younger cohort minimises the risk of contamination; at least in terms of those elements of AR that rely on using the online system (which is most of it).

between groups; typically that data follow a normal distribution (bell curve). Violations of that assumption mean that tests may be inaccurate, or if non-parametric tests are used, less powerful that originally planned. Moreover, they also give rise to questions about what happens if we do or do not find an effect (regardless of the direction of the effect): can we say it is because of AR or is it an artefact of the poor quality data? This is an unsatisfactory position to be in at the end of a trial, particularly one as large as this.

We undertook some further work to understand whether a usable test score could be recovered from the GL-A data. As there was no further information on the psychometric properties of the test provided, we ran a factor analysis to understand if there was any way of using factor scores produced as outcomes. This analysis produced two factors, consistent with the sentence completion and comprehension elements of the test itself. However, there are difficulties with interpreting factor scores, and even more difficulties with translating them into meaningful scores to allow for an assessment of treatment/control differences, and for comparability with other EEF trials. We also asked GL-A whether any weights were applied to questions during scoring that would allow us to re-score data, or otherwise examine this issue further. GL-A reported that they did not use weights in their scoring, adding that the approach to scoring was a simple additive scoring system, with each question weighted equally (so one point for every correct answer). Leaving aside how far educational testing has come in the last few decades, this fact meant we were severely limited in our ability to recover usable data.

Overall, we felt that none of the approaches we could propose would produce data of sufficient quality for the evaluation. Having put these issues to the EEF, we proposed several possible options for the trial, including re-testing all Year 4 pupils. The EEF's grants committee decided to continue funding AR for the target cohorts of pupils for another academic year (2017/18). This meant that the focal analysis and reporting switched from Year 4 pupils to Year 5 pupils, but the inclusion of Year 5 pupils and measuring their outcomes using KS2 data is consistent with the original study design. The continued funding means that target pupils would receive close to six terms of AR overall, which is closer to what the developers believe is required for schools to implement AR. It also means that we will be able to assess the effects of two years' of AR with a one-year gap for the Year 4 pupils when they complete their KS2 assessments.