Statistical Analysis Plan for Read Write Inc. Phonics and Fresh Start

American Institutes for Research

INTERVENTION	Read Write Inc. Phonics and Fresh Start
DEVELOPER	Ruth Miskin Training
EVALUATOR	American Institutes for Research
TRIAL REGISTRATION NUMBER	32235
TRIAL STATISTICIAN	Nisha Rai
TRIAL CHIEF INVESTIGATOR	Nisha Rai and Pooja Nakamura
SAP AUTHOR	Nisha Rai
SAP VERSION	1
SAP VERSION DATE	16 January 2018

Protocol and SAP changes

The only change to the evaluation protocol is the clearer documentation of the various analysis specifications, which we have made in accordance with EEF's guidelines. We also have removed inclusion of geographic controls. As stated in the Protocol, there would be too few areas to reliably estimate variation in effects across areas so we no longer intend to use that control variable. In addition to doing subgroup analysis based on Free School Meal status (as stated in the protocol) we will examine subgroup analysis by gender.

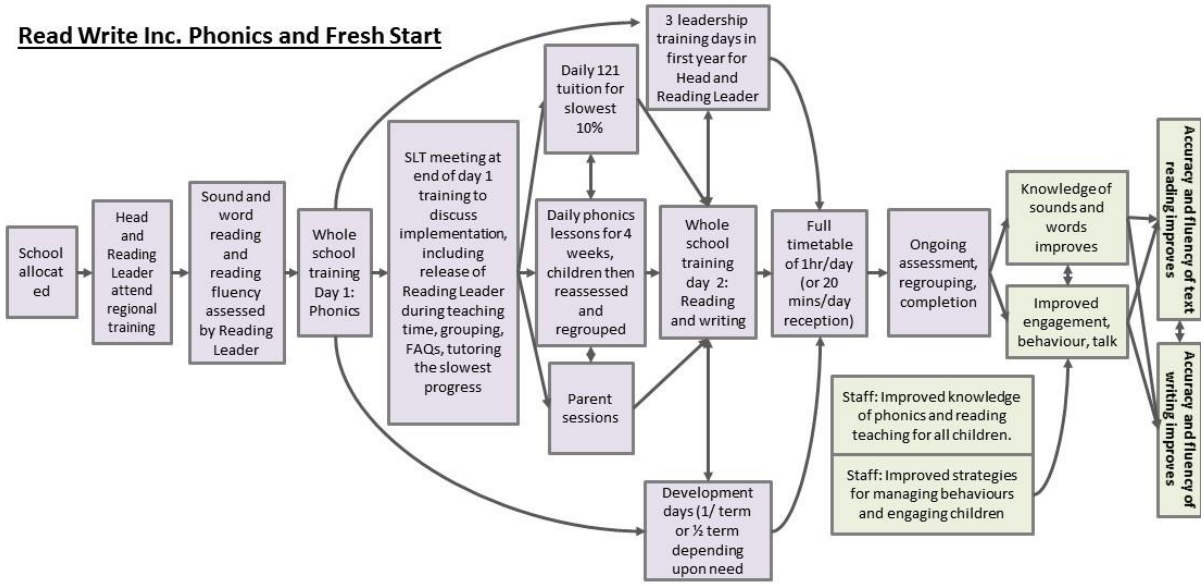
Table of Contents

Introduction

Read Write Inc. Phonics and Fresh Start are phonics-based comprehensive literacy programmes. Read Write Inc. Phonics is a complete literacy programme targeted at 4 to 6 year-olds learning to read and write and 7 to 8 year-olds needing to catch up. Fresh Start is a 33-week catch up literacy programme delivered to target children in Years 5 – 8 identified as having fallen behind in expected standards in reading and writing; it aims to support pupils with reading difficulties at the end of primary school, using systematic phonics approaches embedded in content that is tailored for older pupils. The programme recommends a Fresh Start session every day instead of the usual English lessons and each Fresh Start session should be one hour long.

Both programmes involve the same training (depicted in Figure 1), which is usually delivered to the whole school. The headteacher and reading leader attend a regional two-day training before whole school training. The reading leader assesses the children on sound and word reading and fluency using the programmes' assessment materials. The assessment lasts for approximately five minutes and assesses pupils' phonic knowledge; it is focused on decoding and fluency and does not include a comprehension component. The first day of training is for all teaching staff including the headteacher and focuses on the phonic element of the programmes. Three or four weeks later, the trainer returns to support staff on a normal teaching day. She provides feedback, further training and coaching with a focus on the slowest progress children. The second day of training is approximately four to six weeks after the first training day and focuses on teaching reading and writing. In the first year of implementation the reading leader and headteacher attend two leadership training days. In the second year of implementation, they attend an additional leadership meeting. The trainer continues to visit the school termly to support and coach staff and work closely with the reading leader.

Figure 1. Read Write Inc. Phonics and Fresh Start Training and Support



The Fresh Start programme showed positive results through an EEF aggregated trial, which involved ten schools running their own within-school randomised control trials (RCT). The current study we are conducting is an RCT where randomization was conducted at the level of the school. Thus, while the aggregated trial used a non-standard RCT methodology, this current study serves as an effectiveness trial with implementation conducted at-scale. Furthermore, in addition to exploring results for Fresh Start, this trial will explore results for Read Write Inc. Phonics. With the RCT design we will determine the impact of the programmes on various pertinent reading outcomes.

Study design

Description of population including eligibility criteria

Queens University, Belfast recruited schools that were state funded, with a large (but no more than 50%) share of schools coming from the North East region of the U.K. Additionally, the mean value of the variable proportion of pupils ever having Free School Meal in recruited schools was 30.8%,[1] compared to 14.3% of pupils who were eligible for and claiming free school meals across the country.[2]

At a student level, Read Write Inc. Phonics is delivered to all children in Reception and Year 1 and those who are not meeting national expectations in Years 2 – 4. However, the evaluation will focus on pupils from Reception – Year 2 only. Fresh Start targets eligible pupils in Years 5 – 8 who are below appropriate reading age, but the evaluation will focus on eligible pupils from Years 5 and 6 only. The eligibility criteria for the Fresh Start programme is a school progress measure, through which the children are selected according to their reading levels. Children reading below a reading age of 9.5 years are taught Fresh Start. The programme will be targeted to those children who are not on track to meet national expectations in reading at the end of Key Stage (KS) 2 or those pupils with a reading age below 9.5. While each school uses assessment materials as part of the programme they are using to teach reading, the pupil's progress towards working at National Curriculum Levels is assessed by the teacher.

Description of trial design and arms

The trial is designed as a cluster RCT, and will be used to determine the impact of the programme on various pertinent reading outcomes. The RCT has two arms and will compare the reading achievements of pupils in schools who receive the programmes to pupils in schools who form a two-year wait-list control. The control schools will continue 'business-as-usual' for the teaching of phonics, reading and writing. That is the treatment arm would include those schools who receive Read Write Inc. Phonics and Fresh Start programmes, and the control arm would include those schools who would not initially receive either programme.

Sample size

We determined that a sample of 120 schools split into treatment and control arms in equal proportion would be sufficient for the study. For Read Write Inc. Phonics, the sample would consist of 4,400 Year 2 pupils (37 pupils per school), and for Fresh Start the sample would consist of 1,200 Year 5-6 pupils (10 pupils per school with 5 from each year group). The section Calculation of Sample Size contains information on the actual sample size.

Number and timing of measurement points

For Read Write Inc. Phonics there is one measurement point at the end of the programme. Queens University, Belfast will provide remote oversight to teachers who will electronically administer the New Group Reading Test (NGRT) at the end of the Read Write Inc. Phonics pupil's second year, towards the end of academic year 2017/2018. Teachers will administer the online 13-item adaptive test. Queens University, Belfast will not conduct invigilation for every school but will follow up with schools that are delayed in completing the test. For Fresh Start, the analyses will use Key Stage 2 and Key Stage 1 results from the National Pupil

[1] Based on school statistics Queens University, Belfast collected.
[2] Schools, pupils and their characteristics: January 2016. Department for Education. SFR 20/2016 28 June 2016.

Database (NPD). Queens University, Belfast will have obtained access to the NPD data by November of 2018 for the 2017 – 2018 school year, and analysis will occur within 6 weeks of that date.

Randomisation

We conducted stratified randomisation at the level of the school. The stratification created groups based on the school-level historical KS1 reading score provided by schools as part of the registration form that Queens University, Belfast collected and the government assigned Ofsted grade, a measure of school quality that determines what kind of training they receive. For the historical KS1 reading score, we grouped schools according to quartiles. We assigned the schools into two groupings within the Ofsted grade, those with a high grade of 1 or 2 and those with a grade of 3 or 4.

To facilitate Queens University, Belfast's recruitment of schools, randomisation was conducted in stages, with the first batch of schools in April 2016, the second batch of schools in May 2016, the third batch of schools in June 2016, and the last batch of schools in July 2016. For the second, third, and fourth batches of schools, we assigned quartile levels to KS1 based on the values of the quartiles for the first batch of schools. The analysis of Read Write Inc. Phonics will not use baseline data. The analysis of Fresh start will use KS1 results as a baseline measure of the outcome variable. Since the KS1 assessments occur in school year 2, these will have occurred prior to randomisation, which occurred before the pupils entered year 5.

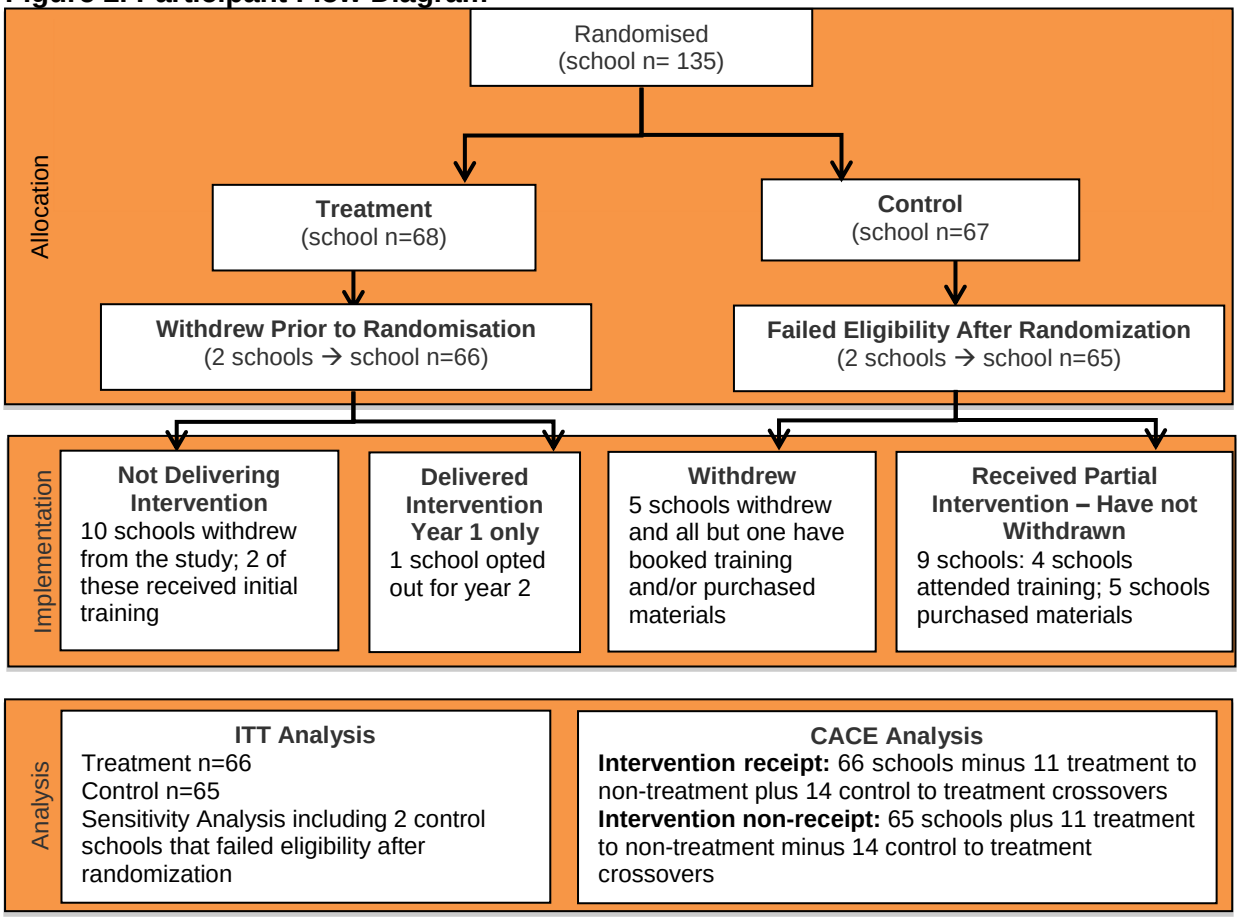
Figure 2 is a participant flow diagram that documents experiences of the schools in the trial. 135 schools were randomised into the 2 groups, with 67 assigned to the control and 68 assigned to the treatment. However, 2 of the schools assigned to the treatment withdrew before they were informed of the results of the randomisation. Thus, we will exclude these 2 schools from all analyses. Additionally, after randomisation results were shared, it was realised that 2 control schools had received the intervention training within the last three years and did not disclose that information in their registration materials. Had this information been properly disclosed they would not have been eligible. For the primary intention to treat analysis, these four schools will be excluded. Thus, for the intention to treat analysis, we will consider 131 schools, 65 in the control group and 66 in the treatment group, which corresponds to the 135 randomised schools minus the 2 treatment that withdrew prior to being informed of their randomisation results and the 2 control schools who were ineligible. We will conduct a sensitivity analysis comparing the primary results with results that include the 2 control schools who were ineligible.

A number of treatment schools did not offer the intervention, and a number of controls schools offered varying degrees of the intervention. Specifically, 10 schools that were randomised to intervention and informed of their results are not delivering the intervention, although two of these schools received the initial training. An additional school received the program in its entirety for the first year but has opted out for the second year.

Additionally, five control schools withdrew from the trial. All of these except for one school have booked intervention training and/or purchased Read Write Inc. materials. Four additional control schools have also attended regional training but have not withdrawn from the study. Five additional control schools have purchased Read Write Inc. Resources since the trial began but have not withdrawn from the study. To reduce the possibility of similar instances

occurring, all control schools have been notified of the importance of maintaining their control status for the research and of the incentive that will be provided at the end of the trial.

Figure 2. Participant Flow Diagram[3]



Because some treatment schools did not offer the treatment and because some control schools offered varying levels of treatment, we will examine the complier average causal effect (CACE), where we will adjust our analyses based on whether the schools actually complied with the assignment. This analysis is described in the section "Non-compliance with intervention" below.

Calculation of sample size

The sample size for the evaluation was calculated using PowerUp![4] based on a 120-school RCT with an even split between the treatment group and the control group. The actual number of schools to be used for the intention-to-treat analysis is 131. In Table 1 we present the power calculations from the initial sample size calculations and those that are based on the actual number of schools in the study.

Table 1. Initial and Actual Power Calculations

[3] Recruitment was led by the Queens University Belfast team.
[4] 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.

	Initial Calculations				Calculations with Actual Sample			
	RWI Phonics		Fresh Start		RWI Phonics		Fresh Start	
Statistic	**(1) ALL**	**(2) FSM**	**(3) ALL**	**(4) FSM**	**(5) ALL**	**(6) FSM**	**(7) ALL**	**(8) FSM**
Alpha Level (α)	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Power (1–β)	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
Rho (Intra-Cluster Correlation - ICC)[5]	0.103	0.103	0.133	0.133	0.103	0.103	0.133	0.133
P (Proportion of Schools Randomised to Treatment)	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
Proportion of variance in outcomes explained by pupil level covariates[6]	0.00	0.00	0.53	0.53	0.00	0.00	0.53	0.53
n (Average Cluster Size)	37	11	10	3	37	9	21	10
J (Sample Size [Number of Clusters])	120	120	120	120	131	131	131	131
N (Total Sample Size)	4,440	1,320	1,200	360	4,832	1,180	2,748	1,354
Minimum Detectable Effect Size (MDES) in SD	**0.184**	**0.222**	**0.215**	**0.267**	**0.176**	**0.222**	**0.193**	**0.205**

Initial Sample Size Calculations

Read Write Inc., Phonics: Assuming a sample of 120 schools with 4,400 Year 2 pupils (37 pupils per school), we conservatively estimated the minimum detectable effect size (MDES) for intent-to-treat estimates of Read Write Inc. Phonics to be 0.184 for all eligible pupils and 0.222 for free school meal (FSM) pupils, assuming 11 FSM pupils per school.

Fresh Start: Assuming a sample of 120 schools with 1,200 Year 5-6 pupils (10 pupils per school with 5 from each year group), we conservatively estimated the MDES for intention-to-treat estimates of Fresh Start to be 0.215 for all eligible pupils and 0.267 for FSM pupils, assuming 3 FSM pupils per school.

Power Calculations with Actual Sample Size

Read Write Inc., Phonics: With a sample of 131 schools, assuming 4,832 Year 2 pupils (37 pupils per school on average), we conservatively estimate the MDES for intent-to-treat estimates of Read Write Inc. Phonics to be 0.176 for all eligible pupils and 0.222 for FSM pupils, assuming 9 FSM pupils per school on average.

Fresh Start[7]: With the sample of 131 schools, assuming 2,748 Year 5-6 pupils (21 pupils – approximately 10 from Year 5 and 10 from Year 6 – per school on average), we conservatively

[5] Intra-Cluster Correlation was informed by the ICCs for KS1 and KS2 for Read Write Inc. Phonics and Fresh Start, respectively. *See* Intra-cluster correlation coefficients as of 24th August 2015 calculated from the 2013 – 2014 National Pupil Data *available at* https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Protocol/ICC_2015.pdf

[6] The proportion of variance in outcomes explained by pupil level covariates was agreed upon during the project's second set up meeting on 5th January 2016. This assumes proportion of variance explained only at the pupil level.

[7] Actual numbers from Fresh Start as reported by Queens University Belfast on August 31, 2017. Note these do not account for fidelity of implementation challenges whereby some schools were not

estimated the MDES for intent-to-treat estimates of Fresh Start to be 0.193 for all eligible pupils and 0.205 for FSM pupils, assuming 10 FSM pupils per school.

Outcome measures

For both programmes we expect to see a positive increase in the primary and secondary outcomes measured, which are described in more detail below. We will use the raw scores for the primary and secondary outcomes, where the raw score is the total number of marks a pupil receives on a test, according to how many questions they answer correctly.

Primary outcome

Read Write Inc., Phonics Primary Outcome: The primary outcome will be New Group Reading Test (NGRT) raw score results at the end of the second year of the programme that will be electronically compiled by Queens University, Belfast.

Fresh Start Primary Outcome: The primary outcome will be KS2 reading fine points raw score from the National Pupil Database for Year 6 eligible students at the end of the first year of programme implementation and Year 5 eligible students at the end of the second year of programme implementation (i.e., the end of Year 6 for students who began programme in Year 5). We will examine this outcome for each cohort separately as well as combined for improved precision.

Secondary outcomes

Read Write Inc., Phonics Secondary outcomes: The secondary outcomes include the teacher assessed KS1 writing at Year 2 and the teacher assessed Year 1 phonics screening check for both years of the programme from the National Pupil Database. We will analyse the data separately for each year of the programme as well as aggregated across years. We will examine both the percentage of children that pass the phonics screening check as well as the raw score of the phonics screening check. Because the teachers will have knowledge of the treatment status when they conduct the Year 2 KS1 Writing and Year 1 Phonics screening check, these outcomes are likely to be more biased than the New Group Reading Test measure. The Department of Education does provide standardized guidelines for the implementation of these tests, which will help minimize bias.

Fresh Start Secondary outcome: The secondary outcome will be the combined KS2 writing results for Years 5 and 6 from the National Pupil Database.

Analysis

Primary intention-to-treat (ITT) analysis

We will analyse the intention-to-treat impact of the programmes using the following difference-in-means specification. All analyses will be ordinary least squares regressions conducted at the individual pupil level using the statistical software package Stata accounting for the nested structure of the data with clustered standard errors.

$$Y_{ijk} = \alpha + \beta R_k + \theta X_{ijk} + \varepsilon_{ijk} \quad (1)$$

Y_{ijk} is the outcome measure for pupil i in classroom j in school k. Read Write Inc. Phonics is administered to all children in Reception and Year 1, so the pupils in the intention-to-treat

implementing Fresh Start. Ruth Miskin Training is currently exploring the reasons for and extent of these challenges.

analysis will be all pupils from the study schools. On the other hand, Fresh Start is administered to eligible pupils, so the pupils in the intention-to-treat analysis will be all pupils identified as eligible by both intervention and control schools prior to randomisation. That is, when analysing the programmes' effects for pupil i in classroom j in school k, Y_{ijk} represents the primary outcomes described in more detail above, specifically for Fresh Start the combined[8] Year 5 and Year 6 KS2 reading results and for Read Write Inc. Phonics the New Group Reading Test at the end of the second year. R_k is a variable that equals one if the pupil's school was assigned to the treatment (which includes both Fresh Start and Read Write Inc. Phonics) and equals zero if not. X_{ijk} is a vector of controls.

For our primary specification, the vector of controls will include the baseline measure of the outcome variable so that we can perform ANCOVA analysis, which will improve our power.[9] The vector of controls will also include the groups created for the stratified randomisation based on the school-level historical KS1 reading score and the Ofsted grades. Due to the randomised treatment assignment, including additional covariates should have no effect on the point estimates of the programmes but may be useful for statistical precision. Thus, we will conduct a secondary specification that will include the following variables from the National Pupil Database (NPD): gender, ethnicity (as defined by the schools), whether the pupil speaks English as a second language, and an indicator for whether the pupil was ever eligible for Free School Meals (FSM), which corresponds to EVERFSM_6_p_ in the National Pupil Database.

In addition to these two specifications, we will report results for the standard EEF model. In this EEF standard model, the vector of controls will include only the baseline measure of the outcome variable. Thus, the covariates for this secondary specification are the treatment group status and baseline measure of the outcome variable only. However, for Read Write Inc. Phonics no baseline measure of the outcome variable exists. For Read Write Inc. Phonics the standard EEF model will include only the covariate of the treatment assignment.

With the ordinary least squares regressions at the individual pupil level, we will use cluster-robust standard errors in our estimates. The use of cluster-robust standard errors accounts for a lack of independence across observations due to the clustering of pupils at the level of school. Doing so ensures that standard errors are properly estimated. β is our coefficient of interest, representing the marginal effect of being part of the Read Write Inc. Phonics or Fresh Start programmes, as compared to the control group.

Interim analyses

No interim quantitative analyses are anticipated.

Imbalance at baseline for analysed groups

In theory, randomization should lead to a balance for outcome and control indicators between the two conditions, but this may not always happen. To assess if the randomisation procedure

[8] As discussed in the primary outcome section for Fresh Start, we will examine this outcome for each cohort separately as well as combined for improved precision. However, the primary outcome will be the combined variable for improved statistical precision.

[9] ANCOVA analysis is not possible with the analysis for Read Write Inc. Phonics since no baseline measure of the outcome variable exists. Since the pre-test is not available for Read Write Inc. Phonics, the secondary specification with additional controls could be very useful for efficiency. In the analysis for Fresh Start, the covariates will include the KS1 reading results as a baseline measures of the outcome variable. It is worth nothing that the KS1 tests are teacher administered so they could contain some bias. However, since the KS1 tests will have been conducted prior to randomisation, the risk of bias is decreased. Nevertheless, for the teacher-assessed KS1 results, we will first normalize the scores to bring all scores into proportion with each other.

created an adequate counterfactual, we will analyse mean differences in the control variables between the treatment group and the control group. Specifically, we will test for balance across ethnicity (as defined by the schools), whether the pupil speaks English as a second language, and whether the pupil was ever eligible for FSM. Our regression will also control for the groups created for stratified randomisation, but as the randomisation specifically balanced within each group we will not test for balance on this dimension. For the analysis of the impact of Fresh Start, we will examine whether the KS1 results are balanced across treatment and control. We will test results using a regression model, which is the same as a t-test but allows for clustering results at the school-level. We will report the means of the control variables and KS1 results for Fresh Start as well as the difference at baseline comparing treatment to control in terms of effect size.

For Fresh Start, we will examine attrition of pupils that were identified as eligible for the programme. Specifically, we will consider the attrition group to be composed of pupils that were identified as eligible for the program, for whom we have KS1 results, but for whom we don't have KS2 reading results. Whereas the panel group will be composed of pupils that were identified as eligible for the program, for whom we have KS1 results as well as KS2 reading results. To see whether the attrition led to an imbalance, we will examine whether the KS1 results and other control variables (age, ethnic group major / minor, language group major/minor, primary and secondary special educational need type, and whether the pupil was ever eligible for FSM) are balanced between those who are in the attrition group and those who are part of the panel.

Missing data

If 5 percent or less of pupils have incomplete information, we will conduct analysis omitting these pupils. In other words, we will conduct analysis using listwise deletion of any pupil with incomplete information. Previous research has found that when 5 percent of the data are missing, bias is low across the various approaches to handling missing data in analysis including listwise deletion.[10]

If more than 5 percent of the pupils have incomplete information, we will first look at the extent of the missing data for each variable to be included in the analysis and the patterns of missingness and consider multiple imputation as an alternative to listwise deletion if necessary. We will conduct t-test comparisons between pupils missing and not-missing data elements. We will also explore a logistic model to explain the patterns of missingness using all available covariates. This descriptive analysis will inform the choice of auxiliary variables to be included in the multiple imputation procedure. Specifically, we will consider the missing group to be composed of Fresh Start pupils that were identified as eligible for the programme, for whom we have missing data. Whereas the non-missing group will be composed of pupils that were identified as eligible for the programme, for whom data are not missing. To see whether missing-ness led to an imbalance, we will examine whether the non-missing variables are balanced between those who are in the missing group and the non-missing group. If any statistically significant differences (at the 5% significance level) are found, multiple imputation with five imputations will be conducted as a sensitivity analysis. We will perform multiple imputation using Stata's command "mi impute mvn," which implements the multivariate normal regression method of imputation. We will use Stata's default number of iterations for the burn-in period, which is 100. We will determine our auxiliary variables through the logistic model

[10] Puma, Michael J., Robert B. Olsen, Stephen H. Bell, and Cristofer Price (2009). *What to Do When Data Are Missing in Group Randomized Controlled Trials* (NCEE 2009-0049). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

described above. In the regressions that use imputed values, we will include a control variable for a dummy that identifies which observations had imputed values.

Treatment Effects in the presence of non-compliance

As described in the randomisation section, 135 schools were randomised into the 2 groups, but we will consider 131 schools for the intention-to-treat analysis. For our primary CACE analysis, we will conduct analyses that consider whether the schools actually complied with the assignment. Of the 131 schools, 69[11] schools received the intervention or an aspect of the intervention and 62[12] schools did not. Our primary CACE analysis considers receipt of the intervention or an aspect of the intervention at the school level as compliance.

Read Write Inc. Phonics is delivered to all children, so pupils who have received the treatment in CACE analysis will be all pupils from the schools that actually received Read Write Inc. Phonics. For Fresh Start, only those pupils who actually receive Fresh Start in any of the schools will be classified as having received treatment in CACE analysis. A number of additional treatment schools failed to deliver Fresh Start and only delivered Read Write Inc. Phonics. Ruth Miskin Training and Queens University, Belfast are working to confirm the exact number of schools/pupils that received Fresh Start, and we will update the Statistical Analysis Plan once those numbers have been confirmed.

We will analyse the average impact of the programmes on compliers using the following two-stage least squares specification. In the first stage depicted by equation (2), we estimate for treatment receipt based on treatment assignment. T_k is a variable that equals one if the school actually received the Read Write Inc. Phonics and Fresh Start programmes.[13] That is, T_k will equal one for the 69 schools that received treatment whether they were allocated to treatment or not. As before, R_k is a variable that equals one if the pupil's school was assigned to the treatment and equals zero if not. X_{ijk} refers to the set of controls for the primary specification in equation (1).

$$T_k = \alpha + \mu R_k + \rho X_{ijk} + \varepsilon_{ijk} \quad (2)$$

The first stage equation enables us to generate predicted values for actual treatment receipt, where the caret symbol indicates the value is estimated.

$$\hat{T}_k = \hat{\alpha} + \hat{\mu} R_k + \hat{\rho} X_{ijk} \quad (3)$$

We can then use the predicted values of treatment receipt $\hat{T}_k$ from equation (3) in our estimate of the complier average causal effect through equation (4).

$$Y_{ijk} = \alpha' + \beta' \hat{T}_k + \theta' X_{ijk} + \varepsilon'_{ijk} \quad (4)$$

Note, actually conducting the two-stage procedure would produce incorrect standard errors. Statistical packages like Stata have built in modules like ivregress, which automatically correct the standard errors. The results of the two-stage least squares regression will provide us with an estimate of the CACE impact, the impact for those schools/pupils who complied with the

[11] Equal to the 66 treatment schools in the ITT analysis minus the 11 treatment to non-treatment crossovers plus the 14 control to treatment crossovers.
[12] Equal to the 65 control schools in the ITT analysis plus the 11 treatment to non-treatment crossovers minus the 14 control to treatment crossovers.
[13] If possible, we will update the definition of actual receipt of treatment through examination of the process data and discussions with the delivery partner and will update the Statistical Analysis Plan accordingly.

programme. We would expect that the CACE impact should be larger (more positive) than the intention-to-treat impact.

In addition to receipt of the interventions at the school level, Ruth Miskin Training has reported varying degrees of compliance with the programmes within schools, particularly with Fresh Start. Ruth Miskin Training is currently exploring the reasons for and extent of the non-compliance. If possible, we will incorporate this information as well as information from the "light-touch" process evaluation as an additional compliance measure in a secondary CACE analysis, and we will update the Statistical Analysis Plan accordingly.

Secondary outcome analyses

We will analyse the intention-to-treat impact of the programmes using the equation (1) as described above. Only now, Y_{ijk} represents the secondary outcome measures for pupil i in classroom j in school k, described in more detail above. Specifically for Fresh Start Y_{ijk} represents the combined Year 5 and Year 6 KS2 writing results, and for Read Write Inc. Phonics Y_{ijk} represents both the New Group Reading Test writing results at the end of the second year and the Year 1 phonics screening check for both years of the programme aggregated. R_k and X_{ijk} remain as described in the primary outcome analysis section.

Additional analyses

Sensitivity Analysis

We will conduct a sensitivity analysis comparing the primary results with results that include the 2 control schools that had received the intervention training within the last three years and did not disclose that information in their registration materials. For the sensitivity analysis, we will analyse the intention-to-treat impact of the programmes using equation (1) as described above but with 133 schools instead of 131.

Exploration of Results from Process Evaluation

Depending on the findings from the process evaluation, we will explore additional analysis and update the SAP accordingly.

Cost-Effectiveness Analysis

We will combine the benefits estimated in our impact analyses with the costs obtained from our implementation research to carry out a cost-effectiveness analysis. Specifically, we will assess the costs of the programmes using the ingredients method. For this purpose, we will need to specify all the ingredients that are necessary to replicate the program and then collect data on the unit costs of all these ingredients (Dhaliwal, Duflo, Glennerster, & Tulloch, 2011). Information on the five major categories of ingredients—(i) personnel and training, (ii) facilities, (iii) equipment and programme materials, (iv) other inputs, and (v) client inputs—need to be obtained. These form the starting point for considering specific ingredients and their costs and are disaggregated into individual cost items (Levin and McEwan, 2001). The approach enables a clear view of how costs are distributed across all types of expenditures. The cost can be in opportunity cost or capital cost, but the ultimate goal is to determine the value of any ingredient.

We will obtain these costs through focus group discussions with school administrators and through the collection of data on actual costs from Ruth Miskin Training. We will then estimate

the costs of the programmes for the average beneficiary and divide these costs by the expected gain in outcome derived from the impact analysis to serve as the cost-effectiveness measure of the programmes.

Subgroup analyses

We will conduct subgroup analysis for the population of FSM pupils.[14] However, because of the small number of eligible FSM pupils per school for Fresh Start we will be unlikely to have the power to detect a separate effect for FSM pupils, although the analysis will still be conducted. We will conduct the subgroup analysis in two ways. First, we will estimate equation (1) only for the sample of FSM pupils. Second, we will examine heterogeneous effects to see if there is a difference between FSM and non-FSM pupil with the equation:

$$Y_{ijk} = \alpha + \beta_1 R_k + \beta_2 R_k * F_{ijk} + \theta X_{ijk} + \delta F_{ijk} + \varepsilon_{ijk} \quad (5)$$

F_{ijk} is an indicator for whether the pupil ever received free school meals. Here the vector of controls X_{ijk} refers to the controls from the primary specification from equation (1), and for the Fresh Start analysis, the KS1 results act as a baseline measures of the outcome variable. With the heterogeneous effects regression, β_1 represents the marginal effect of being part of the Read Write Inc. Phonics or Fresh Start programmes, as compared to the control group, for pupils who do not receive free school meals. β_2 is the marginal effect of being part of the Read Write Inc. Phonics or Fresh Start programmes for pupils who do receive free school meals. This heterogeneous effects analysis will clarify if there is a difference in the programme effect by FSM status.

We will also examine if heterogeneous effects exist along the dimension of gender with the following equation.

$$Y_{ijk} = \alpha + \beta_1' R_k + \beta_2' R_k * G_{ijk} + \theta' X_{ijk} + \delta' G_{ijk} + \varepsilon_{ijk} \quad (6)$$

G_{ijk} is an indicator for whether the pupil is a girl. The vector of controls X_{ijk} is as defined for equation (5). With the heterogeneous effects regression, β_1' represents the marginal effect of being part of the Read Write Inc. Phonics or Fresh Start programmes, as compared to the control group, for boys. β_2' is the marginal effect of being part of the Read Write Inc. Phonics or Fresh Start programmes for girls. This heterogeneous effects analysis will clarify if there is a difference in the programme effect by gender.

Effect size calculation

Results will be reported as Hedge's g effect sizes that standardise the estimated impact using the unadjusted pooled standard deviations of the outcome. That is, we will divide the regression-adjusted impact estimate by the regular unadjusted standard deviations for the full sample included in the analysis, so excluding missing data. This will be done for each of the specifications used.

In addition, we will report the ICC at the level of the school for the outcome measures as well as for the KS1 results used as a baseline control in the Fresh Start regressions.

[14] The FSM variable we use for the subgroup analyses will be the same that we describe above, namely whether the pupil was ever eligible for FSM (EVERFSM_6_p_ in the NPD).

Report tables

In addition to the report tables for the effect size, we will report balance checks through a table that has the variable, mean of control, number in control, mean of treatment, number in treatment, the balance check difference, the balance check standard error, and the balance check p-value.