Study Plan for Ruth Miskin Training's *Read Write Inc.* Phonics

National Foundation for Educational Research

PROJECT TITLE	**Evaluation of Ruth Miskin Training's *Read Write Inc* Phonics**
DEVELOPER (INSTITUTION)	Ruth Miskin Training
EVALUATOR (INSTITUTION)	National Foundation for Educational Research
PRINCIPAL INVESTIGATOR(S)	Simon Rutt
TRIAL (CHIEF) STATISTICIAN	Simon Rutt
STUDY PLAN AUTHOR(S)	Gemma Stone, Afrah Dirie, Simon Rutt

Study Plan version history

VERSION	DATE	REASON FOR REVISION
1.1	09/04/2019	Added information on compliance indicator and analyses
1.0 [*original*]	11/02/2019	

Table of contents

Intervention

WHY; RATIONALE/THEORY/GOAL OF THE INTERVENTION

Read Write Inc. Phonics aims to get children out of the reading gate early regardless of socio-economic status, special need or language status, by providing a whole-school approach to teaching phonics and early reading. Reading is a more important driver of social mobility than socio-economic status (OECD). Children from deprived backgrounds read better and enjoy reading more when they have excellent teachers (Teacher Development Trust and Read On. Get On). Leaders who build a school culture of supportive professional development and teachers who have a love of reading have the biggest impact on children's literacy outcomes.

WHO; RECIPIENTS OF THE INTERVENTION

Four cohorts of primary schools, 72 in total at the beginning of the project, have been recruited to the evaluation by Ruth Miskin Training (RMT). To be eligible, schools had to be 'priority schools', i.e. rated as 'inadequate' or 'requires improvement' by Ofsted and with Phonics Screening Check and Key Stage 1 results below national average, or rated as 'good' by Ofsted but with 'concerning' reading data. 65 of the 72 schools are in the government's priority areas – including the twelve Opportunity Areas. 51 schools are Requires Improvement or below. A target of 70 per cent of schools recruited to be in priority areas was set and exceeded. The four cohorts are shown in table 1 below.

Table 1, cohorts

Cohort	*Start of intervention delivery*	*Year group at start*	*End of intervention delivery*	*Year group at end*
1	Spring 18	Reception (mid)	Spring 20	Year 2
2	Autumn 18	Reception (beginning)	Spring 20	Year 1
3	Autumn 18	Reception (beginning)	Summer 20	Year 1
4	Autumn 18	Year 1 (beginning)	Summer 20	Year 2

The programme consists of whole-school training and ongoing CPD and support over a period of between 19 months and two years depending on the cohort, which is received by all teaching staff, including senior leaders and teaching assistants (TAs). All children learning to read from Reception to Year 2 and any children in Years 3 and 4 not yet reading age-appropriately are taught the *Read Write Inc*. Phonics programme. In small schools, children may be taught in cross-year groups. If there is a school-based nursery attached to the school, children are taught *Read Write Inc*. Phonics in the last term before moving to Reception.

A designated reading leader (RL) is responsible for leading the programme and establishing a coaching cycle of ongoing practice, observation and feedback. The RL role is key for maximising effective implementation. The RL is usually a teacher in Year 1 or 2 (not necessarily the literacy coordinator). The RL does not teach their own group, rather they go

into groups to offer support. They run weekly practice sessions for teaching staff, observe teaching and give feedback. They assess children half-termly and reorganise groupings and set up and monitor daily tutoring to ensure the slowest progress children keep up with their peers.

WHAT; PHYSICAL OR INFORMATIONAL MATERIALS USED IN THE INTERVENTION

Read Write Inc. Phonics teaching resources include handbooks for teachers; handbook for the RL; resources for teaching daily phonic lessons; decodable storybooks including books to read at home; children's writing books; and subscription to online resources through Oxford University Press. There is a RMT online portal for schools, which includes films for RLs to use during weekly practice with their team. The handbooks provide structured teaching guidance, including daily and weekly planning. Teachers are trained on both the programme content and the pedagogical approach – not just what to teach, but how. Consistent behaviour management strategies aim to ensure every child participates in every lesson. Resources are available for parents; this is not an essential part of the programme, but schools might buy some resources and sell them or recommend them to parents. Parent involvement, or purchase of materials, will be not be tracked as part of this evaluation.

WHAT; PROCEDURES, ACTIVITIES AND/OR PROCESSES USED IN THE INTERVENTION

The intervention is a whole-school approach: training is provided for all staff, including the headteacher, teachers, TAs and support staff (although not all staff teach the programme).

Training package:

- Two-day training for RL prior to in-school training (prepares RLs to assess and group students before in-school training, using RMT materials and RWI programme resources)
- Two one-day training days for whole school staff including headteacher and TAs (approximately four weeks between each training day; teaching starts after the first training).
- RL training days – after in-school training: two one-day training sessions for RL along with the headteacher or literacy coordinator. This is focused on assessment, individual tutoring and the coaching cycle
- 16 in-school development days with leadership team and teachers to ensure high quality data-driven teaching. The trainer works more closely with the 20% of children making the slowest progress. They establish a weekly coaching cycle to drive effective teaching. An action plan is completed by the trainer and the reading leader at the end of each development day.
- Schools can also access various free events that are available – visits to model schools, regional meetings.
- Schools are also offered a half-day leadership session with Ruth Miskin.
- They are offered two extra places on phonics training if staff leave and new staff start – if more than two need training schools have to pay (see costs).

TLIF schools receive more intensive support than other schools, who usually receive 6 or 3 development days per year. Schools are offered supply teacher cover for 19 days to release the RL to work alongside the trainer or attend training.

Schools are encouraged to explain the programme to parents and direct them to videos on the website. There are also parental resources that can be purchased if parents choose to (see above – this is not an essential part of the programme). Schools are advised to send storybooks home for children to practise reading after 3-5 days of reading in class. Parents or carers can focus on children's understanding and enjoyment of the stories.

WHO; INTERVENTION PROVIDERS/IMPLEMENTERS

Training and CPD is provided by Ruth Miskin Training consultant trainers. These trainers are well established within the organisation and experienced in providing CPD. The programme is taught by teachers and TAs in programme schools.

HOW; MODE OF DELIVERY

The training and CPD, as set out above, is the method of delivery. Training is for all staff (whole-school approach) but with a focus on RL and headteacher.

WHERE; SETTING/LOCATION OF THE INTERVENTION

Read Write Inc. Phonics lessons are taught in groups in place of literacy lessons. Schools are spread geographically. Most schools are priority schools from government priority areas and 12 Opportunity Areas, identified by the Department for Education (DfE) (see LINK for criteria). Note that the programme is being funded by the DfE's Teaching and Leadership Innovation Fund (TLIF) which supports high-quality professional development for teachers and school leaders in the identified priority and Opportunity Areas.

WHEN AND HOW MUCH; DURATION AND DOSAGE OF THE INTERVENTION

Reception pupils are taught for 20 minutes in term 1, 30 minutes in term 2, and 40 minutes in term 3, then for one hour per day for Year 1 and above. Pupils are assessed every half term and re-grouped (note that as such, their group teacher may change over the duration of the intervention). Lessons include phonics teaching, storybook reading, comprehension, handwriting, spelling, grammar, punctuation and compositional writing. Each storybook is read as part of a 1, 3 or 5 day timetable depending on the length of the text. It is recommended that children take books home after they have been read in class, as described above.

The 20 per cent of children making the slowest progress will receive one-to-one tuition every day, in addition to group teaching, for approximately 5-10 minutes If they are in Reception this should be around five minutes, or ten minutes if they are older. It is particularly important for SEN children, new arrivals and children struggling to keep up in the group setting, who more are likely to fall within the 20 per cent. This is an essential part of the programme.

TAILORING; ADAPTATION OF THE INTERVENTION

The programme is most successful when taught with fidelity and all elements of the programme are taught as per training and the handbook.

COSTS

When signing up to the programme, schools agree to purchase the resources/materials at a 30 per cent discount from OUP. For this intervention, DfE agreed to refund £3,500 to wave one schools if they are not supported by Opportunity Areas leads. This is not currently available for wave two.

Costs for resources range from approximately £5,000 for one form entry school to £8,000 for two form entry school. There is then an ongoing cost of approximately £1,000 per year to re-

purchase consumable items such as writing books. There could be costs for training additional staff if any current staff leave and are replaced. Each school can claim two places for new staff as part of the project. The cost is £260+VAT for each additional member of staff who attends a regional *Read Write Inc* Phonics training. Schools might pay for extra materials; in tandem, they may also elicit some cost savings, for example with reduced time for planning and the lack of need for additional literacy interventions.

Study rationale and background

A number of influential research studies (see Torgerson et al (2006)), attest to the effectiveness of systematic phonics programmes in early literacy teaching. A commitment to the use of systematic synthetic phonics in the teaching of early reading has long been a hallmark of educational policy in England (DfES 2010; Ofsted 2006, Rose 2006). The current national curriculum (published 2013) promotes the view that phonic work isa body of knowledge and skills which all children should be taught, rather than an optional strategy for teaching reading. The phonics screening check, which is a statutory assessment for pupils at the end of Year 1 (implemented 2012), reinforces this approach, as does the matched funding that was provided by the government from 2011 to 2013 for schools to purchase phonics teaching materials aligned with the Department's core criteria for an effective phonics programme.

Underpinning the notion of synthetic phonics is the relationship between sounds and letters in both reading and writing. In English, this relationship is complex. It is necessary to analyse it in terms of phonemes and graphemes, and the correspondences between these. A phoneme is the smallest meaningful unit of sound in a word; a grapheme is a letter or group of letters representing a phoneme. In some cases, the phoneme-grapheme correspondence is simple; for example, the word 'mat' consists of the three phonemes /m/ /a/ /t/, each of which is represented by a single letter. However, in the word 'mash', the single final phoneme is represented by two letters, the grapheme 'sh'. Typically, a single English phoneme can be represented by a number of different graphemes; for example, the phoneme /s/ can be represented by the graphemes 's', 'ss', 'c', 'se' or 'ce'. The essence of a systematic synthetic phonics programme is to teach children these phoneme-grapheme correspondences in a structured way. A high-quality systematic synthetic phonics programme could be expected to take a structured approach to phoneme-grapheme relationships and the order in which they are taught.

This EEF impact study involves the evaluation of a professional development and teaching programme which is based on the *Read Write Inc*. Phonics programme. The *Read Write Inc*. programme itself is on the list of phonics programmes that the Department of Education display on their website. The publishers' self-assessments of the programmes were independently evaluated as meeting the core criteria for an effective phonics programme (although the Department do not rank or endorse the list of programmes).

Theory of Change

Assumptions

- Evidence suggests that reading is a more important driver of social mobility than socio-economic status.
- Children from deprived backgrounds read better and enjoy reading more when they have excellent teachers.
- Leaders who build a school culture of supportive professional development and teachers who have a love of reading have the biggest impact on children's literacy outcomes.
- RMT supports schools teaching the *Read Write Inc* Phonics programme. Schools develop a cohesive approach to teaching phonics and early reading and teach children to decode as the primary strategy for word reading. Word reading is embedded and practiced in closely matched storybooks, enabling children to develop reading fluency which in turn aids comprehension.

Strategies and activities

What is the approach?

- Training for the Reading Leader before whole-school training
- Two whole-school staff training days, with a SLT meeting at the end of the first day
- 16 Support visits to each school, including a Leadership implementation day
- Two Headteacher and Reading Leader Literacy Leaders training days
- Half-termly assessment of children using sound and word reading allowing the reading leader to regroup as necessary
- Daily individual tutoring for the slowest progress twenty percent of children in reception, Year 1 and Year 2.

Resources include:

- Resources include Teacher Handbooks, resources for teaching phonic lessons, decodable storybooks and writing books for children
- Resources are available for schools to purchase from Oxford University Press

Target groups

Schools: Primary

Regions: Priority Areas (as set by DfE)

Pupils: All pupils in Reception, Year 1 and Year 2

Short-term outcomes (1-2 years)

Pupil impact:
Primary outcome:
Improved phonics attainment, measured by the Phonics Screening Check.

Secondary outcome:
Improved reading attainment, measured by Key Stage 1 reading outcomes.

School-level impacts:

- Teacher confidence in teaching phonics and early reading improves
- Reading Leader has enhanced ability to lead literacy in school, including practices such as cycle of practice, coaching and feedback

Longer-term outcomes (2-5 years)

- Improved attainment in reading, at KS2
- Improved attainment in writing, at KS2

Impact evaluation

Research questions

Primary Question

RQ1: What is the impact of the Ruth Miskin Training: *Read Write Inc*. intervention on the overall phonics ability of children aged 5-6 years old?

Secondary Questions

RQ2: What is the impact of the Ruth MiskinTraining: *Read Write Inc*. intervention on the overall reading ability of children aged 6-7 years old?

RQ3a: Are effects on phonic knowledge different for pupils eligible for free school meals (FSM)? If so, how?

RQ3b: Are effects of the intervention on reading ability different for pupils eligible for FSM? If so, how?

RQ4: Is there an interaction between fidelity and attainment for schools who have received the intervention?

Design overview

Design type		Difference-in-Differences with matched schools
Unit of analysis (school, pupils)		Pupils
Number of Units (Treatment, Comparison)		Treatment: 5822 pupils Comparison: 5822 pupils Total: 11644 pupils
Outcomes	primary	Phonics Screening Check (PSC) performance
	secondary	Key Stage One (KS1) reading performance PSC performance as binary measure
Outcome sources (instruments, datasets)	primary	PSC raw scores collected from schools.
	secondary	KS1 reading raw scores collected from schools PSC outcomes collected from schools.

Participants

The school population is maintained primary schools in England with Key Stage 1 (KS1) pupils. The eligibility criteria, agreed with the Teaching and Leadership Innovation Fund (TLIF) were:

- At least 70% of schools must be from an opportunity area or priority area

- Schools should have an Ofsted rating of 'Require Improvement' or 'Inadequate'. 'Good' schools with literacy data that is causing concern are considered on a case by case basis if in an Opportunity Area or Priority Area ('Outstanding' schools not considered)
- Schools must not have had in-school training or support from Ruth Miskin Training in the last two years

62% of the intervention sample are Ofsted rated 'Requires Improvement' or 'Inadequate' and from an opportunity area or priority area. The other 38% represent two other combinations:

1) Rated 'Requires Improvement' and below, and outside of the opportunity areas and priority areas

2) Ofsted rated 'Good' but have literacy data that is causing concern and are in an opportunity area of priority area.

These are schools that are performing well below average, based on previous cohorts' Phonics Screening Check and KS1 reading performance.

The evaluation aim is to assess the impact of the programme on pupil's phonics and reading attainment; participants are KS1 pupils in Reception to Year 2. The sample size for treatment schools is set as 71 primary schools with pupils in KS1, split across two waves and four cohorts.

1) Wave one (35 schools) consists of pupils in cohort 1 that will participate in the programme from the middle of Reception in Spring 2018 to the end of Year 2 in March 2020[1] and pupils in cohort 2 that will participate in the programme from the beginning of Reception in Autumn 2018 to the end of Year 1 in March 2020[2].
2) Wave two (36 schools) consists of pupils in cohort 3 that will participate in the programme from the beginning of Reception in Autumn 2018 to the end of Year 1 in March 2020[3] and pupils in cohort 4 that will participate in the programme from the beginning of Year 1 in Autumn 2018 to the end of Year 2 in March 2020[4].

To date, in Wave 1 one school closed, but they will not be replaced. After all wave 2 schools have been recruited, it has been agreed with DfE that if any school drops out after training has occurred they will not be replaced. However, if dropout occurs before September 2018, RMT will recruit an additional school.

We will only select pupils from these cohorts who were on roll at the school at the time of the initial/whole school training, or joined up to one month after the training. Late arrivals will be excluded from analysis.

Study design

This evaluation will use a difference-in-differences (diff-in-diffs) with matched schools approach, which aims to assess the impact of the programme on pupil's reading ability and phonic knowledge. An identified assumption of this design is the parallel trends assumption, which implies that in the absence of the intervention, the treatment and comparison groups

[1] The funded project expires in March 2020.
[2] The funded project expires in March 2020.
[3] The funded project expires in March 2020.
[4] The funded project expires in March 2020.

will follow the same trend (Liu et al, 2010). Comparison schools will be selected by matching eligible schools based on observable characteristics by propensity score matching (PSM) with details on how this is carried out provided below.

The primary analysis will use pupils in cohorts 2 and 3 to analyse the impact of the Ruth Miskin Training: *Read Write Inc* intervention on the overall phonic knowledge.

We intend to carry out two secondary analyses:

i. A pupil-level analysis will use pupils in cohorts 1 and 4 to analyse the impact of the Ruth Miskin Training: *Read Write Inc* intervention on the overall reading ability.

ii. A pupil diff-in-diffs analysis will use pupils in cohorts 2 and 3 to analyse the impact of the Ruth Miskin Training: *Read Write Inc* intervention on the overall phonic knowledge[5].

More detail on the analyses to be undertaken provided below.

Outcomes and other data

PRIMARY OUTCOME

The primary analysis will be to estimate the programme's impact on phonics performance at the end of Year 1. We will measure the primary outcome using the PSC test administered at the end of Year 1. NFER will ask all schools to provide PSC raw score data for Year 1 pupils in the Summer 2020. We will also require previous cohorts (2015-2019) PSC raw scores, anonymised, from the full NPD[6]. We will use these scores in the primary analysis to see the trajectory of PSC raw scores and the treatment effect of the intervention.

SECONDARY OUTCOMES

We intend to carry out two secondary outcome analyses:

i. The first analysis will be to estimate the programme's impact on reading performance at KS1. We will measure the secondary outcome using pupil performance measures from statutory assessment data. NFER will ask all schools to provide KS1 reading raw score data for Year 2 pupils in Summer 2020. This is because the performance data released on the NPD is only for age-related expectations, which we consider not to be sensitive enough for this type of analysis. We will collect from the NPD 2017 KS1 reading outcomes[7], anonymised to control for the school level attainment as mentioned in the analysis section.

ii. The second analysis will be to estimate the programme's impact on phonics performance at the end of Year 1. We will measure the secondary outcome using the PSC test administered at the end of Year 1. We will convert the primary outcome into a binary variable[8] (*phonics outcome)* which measures the phonics outcome of

[5] We will convert the primary outcome measure into a binary variable which will become the outcome variable for this analysis.
[6] This is reliant on being granted access to the data released in October 2020 that would contain outcomes from the phonics assessment. The NPD data item reference for phonics raw scores is *PHONICS_MARK.*
[7] This is reliant on being granted access to the data released in October 2020 that would contain outcomes from the KS1 assessment. The NPD data item reference for KS1 reading outcomes is *KS1_READ_OUTCOME.*
[8] RM requested this as the audience might understand the interpretation of the analysis using a binary measure better than using a continuous measure.

pupils[9]. We will also require previous cohorts (2015-2019) phonics outcomes, anonymised, from the full NPD[10]. We will use these scores in the analysis to see the trajectory of phonics outcomes and the treatment effect of the intervention.

OTHER DATA

NFER will collect Unique Pupil Numbers (UPNs) for all participating pupils, together with the names, Teacher Reference Numbers (TRNs) and contact details of participating reading leaders, school leaders and teachers. We have costed to send a brief proforma to schools taking part in RMT to collect this information.

We will produce a dataset, which contains information that we will collect directly from schools in January 2020. We will collect information about the school (Unique Reference Number (URN), school name, region, school size) and their pupils (UPN and date of births). We will also collect from schools, in summer 2020, their 2020 PSC (for Year 1 pupils) and KS1 reading (for Year 2 pupils) raw scores. If we have missing school level data , we will consider using other sources like Edubase and draw on MI data collected by the DfE[11] (depending on its completeness, quality and access permissions) to replace this missing data.

We will send this dataset to NPD to match for each pupil their start date, gender, ethnicity, SEN level and whether they are eligible for FSM.

Additionally for the primary outcome, we will additionally request pupil level anonymised PSC raw scores between the years 2015-2019. We will use this data in the primary outcome diff-in-diff analysis.

For secondary outcome (i), we will additionally request pupil level anonymised KS1 reading outcomes for 2017. We will be aggregate this data to create a school level measure of KS1 performance, which will be used in the secondary outcome analysis. For secondary outcome (ii), we will additionally request pupil level anonymised PSC outcomes between the years 2015-2019. We will use this data in the secondary outcome diff-in-diff analysis.

Sample size calculations

The primary research question is intended to focus on all pupils for the primary analysis, but research questions RQ3a and RQ3b will focus on FSM-eligible pupils for sub-group analyses.

The trial consists of 142 schools in total, 71 in each study group (treatment and control). With the assumed average of 82 pupils per school and 142 schools, the overall sample is 11644 pupils.

The average percentage of eligible FSM was obtained by using NFER's Register of Schools (ROS, a database containing details of all schools). Considering schools in the treatment

[9] Using the threshold mark released by DfE in Spring 2020, a *phonics outcome* of *Wa* indicates that a pupil has met the expected phonic decoding standard for a pupil at the end of Year One whilst a *phonics outcome* of *Wt* indicates that a pupil has not met the expected phonic decoding standard.

[10] This is reliant on being granted access to the data released in October 2020 that would contain outcomes from the phonics assessment. The NPD data item reference for phonics outcome is *PHONICS_OUTCOME.*

[11] We understand that DfE is collecting MI data for all TLIF projects.

group only, NFER matched these schools to ROS by URN to obtain percentages of pupils eligible for FSM. This led to a pupil eligible FSM rate of 25.17% and based on the average number of pupils per class (82), the expected number of eligible FSM pupils per class is 21. Thus, the overall sample for FSM eligible pupils in 141 schools is 2982.

Without a writing trial pilot using our chosen assessment regime, parameters for sample size calculations must be estimated using comparable studies. As all evaluations in the EEF report archive that have phonics or KS1 literacy as a primary outcome involved randomised designs, NFER has considered ICC values of suitable studies. The improving numeracy and literacy evaluation (Education Endowment Foundation, 2015) with the outcome as Key Stage 1 literacy suggests a value of 0.09 for all pupils and 0.11 for FSM eligible pupils. Our evaluation does not require the calculation of a pre-test/post-test correlation as there is no baseline measure. We also assumed 80% power and alpha at 5%.

All sample size calculations were carried out using a purpose-built Excel spreadsheet.

The table below gives the MDES calculations for all pupils and FSM eligible pupils. For the primary analysis of all pupils, the MDES was estimated to be 0.16 and for sub-group analyses of eligible FSM pupils, the MDES was estimated to be 0.19.

		Study Plan	
		OVERALL	FSM
MDES		0.16	0.19
Pre-test/ post-test correlations	level 1 (pupil)	NA	NA
	level 2 (class)	NA	NA
	level 3 (school)	NA	NA
Intracluster correlations (ICCs)	level 3 (school)	0.09	0.11
Alpha		0.05	0.05
Power		0.8	0.8
One-sided or two-sided?		Two	Two
Average cluster size		82	21
Number of schools	treatment	71	71
	control	71	71
	total	142	142
Number of pupils	treatment	5822	1491
	control	5822	1491
	total	11644	2982

Selection of the comparison group

If we applied the same selection criteria as TLIF to choose a suitable pool of comparison schools, 2031 schools would be eligible for sampling, which we believe may include those in the treatment group. The table below shows comparisons between the intervention sample

and our pool of potential comparison group in terms of the selection criteria provided by TLIF.

Selection Criteria	Intervention sample	Pool of potential comparison schools
OA/PA and Ofsted 'Requires Improvement' or 'Inadequate'	44	279
Not OA/PA and Ofsted 'Requires Improvement' or 'Inadequate'	7	375
Not OA/PA and Ofsted 'Good' and KS2 reading below national average	20	1377

Using data provided by EEF/Ruth Miskin Training on the 71 treatment schools, we plan to produce a sample of similar schools, which we will recruit into a comparison group. We will select the comparison group of 71 schools (provided that these schools have not received RMT) by matching eligible schools to the observable characteristics of the treatment schools.

We will use freely-available school-level data to identify relevant characteristics. We will begin by using characteristics relating to the selection criteria. These are historic performance data from before the start of programme delivery (2017 performance for Cohort 1 and 2018 performance for all other cohorts), OA/PA (a binary variable that will be created with $0 = \text{not OA/PA}$ and $1 = \text{OA/PA}$) and Ofsted rating. We will then carry out the matching process described below and assess the matching quality of observable characteristic balance in the matched sample (with how this is carried out mentioned below). If there appears to be an imbalance (a statistically significance difference between groups), we will consider using the following observable characteristics of the treatment schools until we obtain a matched dataset[12]:

- region (North or South)
- school size (as a continuous variable)
- gender
- school FSM

According to Little (2011), there are three main decisions affecting a matched dataset; the choice of measuring distance, the choice of matching strategy, and choice of algorithm to perform matching.

There are many different ways of measuring distance $(\mathrm{D_{tc}})$ between the observable characteristics of study groups, the most common are:

1) Exact[13]:
   - $\mathrm{D_{tc}} = 0$ if $\mathbf{X_t} = \mathbf{X_c}$
   - $\mathrm{D_{tc}} = \infty$ if $\mathbf{X_t} \neq \mathbf{X_c}$
2) Mahalanobis[14]:
   - $\mathrm{D_{tc}} = \sqrt{(\mathbf{X_t} - \mathbf{X_c})' \, \mathbf{S_X}^{-1} (\mathbf{X_t} - \mathbf{X_c})}$
3) Propensity score[15]:

[12] We will also consider using additional years for historic performance data.
[13] $\mathbf{X}_t$ is a vector of observable characteristic values for the treatment group and $\mathbf{X_c}$ for the control group
[14] $\mathbf{S}^{-1}$ is the covariance matrix of the observations
[15] π_t is the probability of belonging in the treatment group , given the observable characteristics and π_c is the probability of belonging in the control group

- $\mathrm{D_{tc}}(\mathbf{X_t}, \mathbf{X_c}) = |\pi_t - \pi_c|$

The exact method is the most straightforward way but, it is not ideal in our case as we have some continuous observable characteristics and it is unlikely that the value for these covariates is exactly the same for both study groups. An extension of exact matching is coarsened exact matching (CEM), which allows continuous or ordinal data to be segmented into strata. However, if the strata are too complex, this will result in poor matches as CEM requires an exact match ($\mathrm{D_{tc}} = 0$).

The Mahalanobis method is not ideal when the vector of observable characteristics ($\mathbf{X}$) is highly dimensional[16] as it does not take into account all interactions between covariates. In our case dimensionality is unlikely to be an issue however, our observable characteristics does include several dichotomous variables (i.e. region) and the Mahalanobis method may not be the most suitable method for such variables. Using propensity scores overcomes this through collapsing the vector of observable characteristics into a scalar propensity score.

A propensity score is the probability of participating in a given intervention, given a set of observable characteristics. Propensity scores are estimated using a logistic regression model; the outcome of interest in the estimation of propensity scores is the binary indicator of whether the pupil is part of the treatment group.

We have chosen to use a one-to-five matching strategy with replacement to avoid the issue that arises with one-to-one matching in terms of response rates. By using a one-to-one matching strategy, we assume a low response rate and so increasing the number of schools that a treatment school can be matched to will help us obtain a comparison group of 71 schools. We do not want the nearest propensity scores to be too far away resulting in less accurate treatment effects. A solution to this is to impose a maximum tolerance on the distance between the propensity score of a treatment school, and the propensity score of its matches known as a caliper. This prevents residual bias caused by poor matching. We will use a caliper of 0.25 as recommended by Rosenbaum and Rubin, 1985.

The last decision to affect a matched dataset is the type of matching algorithm used. As our matching strategy involves replacement, the only matching algorithm suitable is nearest neighbour matching. This assigns a set of nearest propensity scores (neighbours) to a treatment school. Since each treatment school is matched based on a minimum distance between its propensity score and the score of its nearest neighbours, the overall heterogeneity of the matched dataset is reduced.

We will be computing propensity scores as well as creating a matched dataset using the MatchIt (Ho et al., 2013) package in R (R Core Team, 2017). Once a matched sample has been formed, the diff-in-diffs treatment effect can be estimated by comparing the outcomes between treatment schools and comparison schools through the use of multi-level regression models to answer the research questions described above.

We aim to produce a sample of comparison schools in Autumn 2019. Once we have recruited these comparison schools, we will collect their phonics and KS1 reading raw scores and outcomes at the end of the evaluation period (Summer 2020).

A senior researcher from NFER's Research and Product Operations (RPO) team will lead the recruitment of the comparison group schools. We will design carefully worded recruitment documents (Memorandum of Understanding (MOU), project information sheets

[16] High dimensionality refers to datasets where the number of covariates exceeds the number of observations

to ensure that schools know what they are signing up to, and a reply form), which will clearly and concisely outline the purpose of the research and the value of participating. On receipt of schools' recruitment documentations, our specialist Telephone Unit will call the school to confirm their participation and answer any questions they might have. In line with NFER's Code of Practice and Working with Schools Policy, our processes will focus on reducing the burden on schools taking part and we will support the schools throughout the evaluation period.

As Ruth Miskin Training has already recruited intervention schools, NFER will carefully manage the relationship with them, as well as recruiting and maintaining a relationship with comparison schools. This will involve regular communications and updates via email and post, alongside a clear memorandum detailing the schools' responsibilities at outset. As the intervention schools were recruited without NFER's involvement, it will be crucial to ensure that RM's and NFER's messaging is consistent and well integrated. This will be managed by a shared communications plan and RM checking documents before they are sent to schools to ensure consistency in terminology.

Imbalance between groups

Before we implement a matching method to obtain a dataset that includes the 71 treatment schools and 71 matched control schools, we need to examine any difference between treatment and potential control schools (2031 schools) for the primary outcome variable and observable characteristics. We will identify any balances or imbalances for continuous variables through comparing means and conducting t-tests to test if there is a statistically significant difference between study groups. For binary variables, we will identify imbalances by computing cross tabulations and conducting chi-square tests of homogeneity. As the matching method has not been implemented, we assume that there will be an imbalance between study groups, which is identified through obtaining a *p*-value less than 0.05 indicating that there is a statistically significant difference between the study groups for each observable characteristic. We will also present the standardised difference for each of the variables in terms of effect size.

After executing a matching algorithm and obtaining a matched dataset, we will assess the matching quality of observable characteristic balance in the matched sample. We will visually inspect any imbalance by plotting the mean of each observable characteristic against the estimated propensity score. The treatment and control groups will have (near) identical means of each observable characteristic at each value of the propensity score, if matching is successful. We will also identify any balances or imbalances by performing the same tests as carried out prior to matching. We assume that there will be a balance between study groups, which is identified through obtaining a *p*-value greater than 0.05 indicating that there is not a statistically significant difference between the study groups for each observable characteristic. We will also present the standardised difference for each of the variables in terms of effect size.

We will present the differences in covariates between the intervention schools and the unmatched pool of comparison schools, and then in comparison with the matched comparison schools to demonstrate how propensity score matching improves balance in observable characteristics.

We do not envisage that these tests indicate an imbalance between the study groups after matching. However, if there are any imbalances in observable characteristic values between

study groups, we will consider using a different matching method (using a one-to-many matching strategy without replacement and an optimal matching algorithm as an example) and assess the matching quality of this matched dataset using the same method outlined above. If this matched dataset also indicates imbalance between study groups, other different matching methods are considered until we obtain a matched balanced dataset.

Once we obtain a matched dataset and after assessing the matching quality indicated that the study groups are balanced in terms of observable characteristics, we will assess the balance of this sample relating to missing data. To see whether attrition leads to an imbalance, we will examine whether observable characteristics are balanced between two groups of pupils; the attrition group (pupils for whom we have observable characteristic values, but we do not have raw scores of our outcomes) and our non-attrition group (pupils for whom we have observable characteristics and PSC raw scores).

Any imbalance between these groups for categorical variables will be identified using cross tabulations and a chi-square test of homogeneity will be carried out to test if there is a significant difference between study groups. These tests are carried out at pupil level and the categorical variables considered for analysis are:

- gender (which will be tested at pupil level)
- year group
- region
- area
- ethnicity
- SEN level
- pupil FSM

Any imbalance between the two groups for continuous variables can be assessed by calculating the mean of each variable for each study group and the standardised differences in terms of effect size. For such variables, two-independent samples t-test is carried out to test if there is a significant difference between study groups. These tests are carried out at pupil level and the continuous variables considered for analysis are:

- school size
- school FSM
- historic performance (2017 performance for Cohort 1 and 2018 performance for all other cohorts)

If our two groups are not equivalent (i.e. statistically significantly difference on any observable characteristic), we will identify the likely missingness mechanism and based on this outcome, we will consider the appropriate technique for handling missing data. More details on this mentioned in the missing data section below.

Primary analysis

The primary analysis will be a pupil diff-in-diffs after matching treatment and comparison schools as described above. A multilevel model with two levels (school and pupil) will be used for the analysis to account for clustering. The main analysis will use cohorts 2 and 3 to assess the programme's impact on phonics. This will be determined by fitting a model with a dependent variable, using the primary outcome described above.

We will request, from NPD, anonymised Phonics Screening Check (PSC) raw scores for all pupils in treatment and comparison schools for the years 2015 to 2019. We will use this data to see how the schools are progressing each year, but the main change we would be looking for is between 2018 and 2020. We will include an ordinal variable *year* with 2018 as the reference group into the regression model to measure the rate of change on the PSC raw scores with year whilst holding all other covariates in the regression model constant.

The regression model is given by

$$Y_{it} = \beta_0 + \beta_1 \text{treatment}_i + \boldsymbol{\beta_2} \text{year}_t + \boldsymbol{\beta_3} \text{treatment}_i * \text{year}_t + \boldsymbol{X}'_{it}\boldsymbol{\beta} + \epsilon_{it}$$

where Y_{it} is the raw phonics raw score of a pupil in treatment i at year t, treatment_i is a dummy variable set to 1 if the pupil is in the treatment group at any year t, year_t is an ordinal variable (from 2015 to 2020) with 2018 as the reference group.

The vectors of coefficients $\boldsymbol{\beta_2}$ and $\boldsymbol{\beta_3}$ represent the change in phonics attainment for each year compared to 2018 and the change in phonic attainment for the interaction between the treatment group and each year. The diff-in-diffs estimate of the treatment effect represented by the coefficient that explains the relationship between treatment given and the year 2020.

Other covariates for this model ($\boldsymbol{X}'_{it}$) include: *gender* (male as the reference group), *region* (North as the reference group), *ethnicity* (white British as the reference group), *SEN level* (1 as the reference group), and *pupil FSM* as binary variables, *school size* and *school FSM* as continuous variables.

We have included such covariates as it is likely that there will still be some variance left after selecting the comparison group of 71 schools by matching eligible schools to the observable characteristics of the treatment schools using PSM.

The error associated with the i^{th} pupil at year t is given by ϵ_{it} and this model will be run in R using the lme4 (Bates et al., 2015) package with a full syntax trail.

Further analyses

SECONDARY OUTCOME ANALYSES

We intend to carry out two secondary outcome analyses. The first analysis will answer RQ2 through carrying out a pupil-level analysis after matching treatment and comparison schools as described above. A multilevel model with two levels (school and pupil) used to account for clustering. We will perform this analysis on cohorts 1 and 4 and will identify differences in KS1 reading attainment between the treatment and comparison group, whilst controlling for school level prior attainment and other pupil background and school level factors. The prior attainment variable will control for pupils being in high and low performing schools. We will collect from the NPD 2017 KS1 reading outcomes at pupil level (anonymised) to use as a prior school-level measure. We will aggregate these scores to school level to obtain the percentage of pupils that met the expected standard in the test.

This will be determined by fitting a model with a dependent variable as KS1 reading attainment as measured by secondary outcome (i) described above.

The regression model is given by

$$Y_i = \beta_0 + \beta_1 \text{treatment}_i + \beta_2 \text{prior} + \beta_3 \text{cohort}_i * \text{treatment}_i * \text{prior} + \boldsymbol{X}'_i\boldsymbol{\beta} + \epsilon_i$$

where Y_i is the raw KS1 reading raw score of a pupil in treatment i, $\mathrm{treatment}_i$ is a binary variable set to 1 if the pupil is in the treatment group, prior is a continuous[17] variable measuring the prior attainment, and cohort_i is a binary variable which identifies which cohort a pupil is in[18]. The treatment estimate of the treatment effect is given by β_3.

Other covariates for this model ($\boldsymbol{X}_i'$) include: *gender* (male as the reference group), *region* (North as the reference group), *ethnicity* (white British as the reference group, *SEN level* (1 as the reference group), and *pupil FSM* as dichotomous variables, *school size* and *school FSM* as continuous variables.

We have included such covariates, as it is likely that there will still be some variance left after selecting the comparison group of 71 schools by matching eligible schools to the observable characteristics of the treatment schools using PSM.

The error associated with the i^{th} pupil is given by ϵ_i and we will run this model using the lme4 package in R with a full syntax trail.

The second analysis will be a pupil diff-in-diffs after matching treatment and comparison schools as described above. This will be similar to the primary analysis except the dependent variable is the secondary outcome (ii) described above and the ordinal variable *year* will be the anonymised phonics outcome of pupils in the intervention and control schools between the years 2015 and 2019 (requested from NPD). We will run this model using the lme4 package in R with a full syntax trail.

SUBGROUP ANALYSIS

The primary outcome model will be modified for the following subgroup analysis specified in the study plan: FSM-eligible pupils (we will use the FSM indicator as from the NPD) and gender. We will carry out this analysis through the use of interaction terms in the model (i.e. the product of the variable, year and treatment group).

Both secondary outcome models will be modified for the following subgroup analysis specified in the protocol: FSM-eligible pupils (we will use the FSM indicator as from the NPD). We will carry out this analysis through the use of an interaction term in the model (i.e. the product of the variable and treatment group). We will also modify the model for the analysis of secondary outcome (ii) to include an interaction term (the product of gender, year, and treatment group).

TREATMENT EFFECTS NON-COMPLIANCE

The main analysis will be followed by a CACE (Compliance Average Causal Effect) in order to assess the impact of non-compliance on the outcome measure. The information on the number of training and development days will be collected by Ruth Miskin as agreed with NFER.

These measures will be included in the analysis:

[17] This is dependent on data for prior being normally distributed. If not, we may consider categorising prior into quintiles.

[18] Pupils in cohort 1 will see an effect of the intervention earlier compared to cohort 4 as they will participate in the programme in Spring 2018 whereas cohort 4 will begin in Autumn 2018.

Level of compliance	Description
High	Completed the in-school training day Completed all 16 in-school development days
Medium	Completed the in-school training day Completed 1-15 in-school development days
Low	Completed the in-school training day Not completed any in-school development days
Zero	Any schools not completing any training

Schools may potentially have unobserved characteristics that have an influence on both the compliance with the intervention and academic attainment. Therefore, a two stage least squares model will be used to calculate the CACE estimate (Angrist and Imbens, 1995). The first stage of the model will be compliance regressed on all covariates used in the analysis of the primary outcome and in addition, will include a binary variable that indicates a pupil's pre-treatment allocation (as an instrument variable). The second stage of the model will be the primary outcome regressed on covariates used in the main model as well as the predicted values of compliance obtained from the first stage of the model. The coefficient of the compliance variable in the second stage of the model is the CACE estimate of the compliance effect.

In the event that there are no confounding factors affecting compliance and attainment, the CACE estimate will be equal to an intention-to-treat estimate. We will use the ivpack package (ivreg function) in R to perform the CACE analysis on the primary outcome only.

MISSING DATA

Prior to commencing any analysis we will look at the amount of missing data, but as all the data is coming from statutory testing we do not envisage this being problematic. If there is less than 5% missing data, we will consider carrying out a complete cases analysis as this is unlikely to be biased.

However, if there is more than 5% missing data, following EEF guidelines, the extent of missingness and the pattern of missingness are important when analysing missing data. The number of complete cases (a complete case would be a case with complete data on all variables of interest) will be reported.

The extent of missingness will be also quantified and reported. Based on these numbers we will discuss the adverse effects of the missing data on sample size and thereby on statistical power and other implications.

We will identify likely missingness mechanisms; Missing completely at random (MCAR), Missing not at random (MNAR), and Missing at Random (MAR)[19]. We assume that the most likely scenario for missing endpoint data is MNAR. This is because as our endpoint measurement are 2020 PSC and KS1 reading raw scores, pupils with missing data could depend on variables that are not observed i.e. Less able pupils are more likely to have missing data on the dependent variable and this factor (pupil's ability) was not measured.

To test these assumptions we will conduct diagnostics to establish any measurable predictors of withdrawal from the study. Initially, we will look for any imbalances between the groups (attrition and non-attrition) with description on how we intend to carry this out mentioned above. If our groups are not equivalent (i.e. statistically significantly difference on

[19] For definitions see https://www.ncbi.nlm.nih.gov/books/NBK493614/

any measure, carrying out a 'complete case analysis' (using only cases with complete data) may be biased as the study groups are not representative of the original sample. This analysis also shows us whether cases with particular characteristics are more likely to have dropped out (biased attrition).

We will also run logistic regression models with missingness on each variable of interest (1=missing, 0=otherwise) as the dependent variable and observable characteristics as independent variables. Tests will be conducted at both the student and school levels, as there are relevant school-level variables available.

We will carry out Little's MCAR test (McKnight et al. 2007 pp.93-94) which will be performed in R using the package BaylorEdPsych (Beaujean, 2012) as we can examine multiple variables simultaneously. The null hypothesis for Little's MCAR test is that the data is MCAR, and significance values of 0.05 suggest the data is likely to be MAR or MNAR.

Given the need to compare estimates of the intervention's effect under different missing data assumptions, in order to generate sensible estimates (although less precise) under the MCAR assumption, our first step will be to perform a complete case analysis for each model.

If, for each model, only the endpoint outcome is missing, we can then add any covariates predictive of non-response to this model and thereby produce valid estimates under the MAR assumption. If we have MAR outcome data and covariates at endpoint, the best strategy is multiple imputation which will produce estimates under an MAR assumption.

Multiple imputation uses complete cases to make multiple estimates of each missing value which are then used to make a single best estimate. The variables used are generally those to be included in the model and those which are associated with missingness. We will use the mice (Buuren and Groothuis-Oudshoorn, 2011) package in R, with predictive mean matching as our imputation method.

Sensitivity analysis will then be used to test the sensitivity of these results to the possibility that the data are missing not at random (MNAR).

We will use sensitivity analysis to assess the robustness of the analysis we have conducted under the MAR missing data assumption via multiple imputation. We will consider imputing 'extreme' values to the missing cases for the relevant variables (such as all 0s or all 1s for a binary variable) and seeing how, if at all, this changes the conclusions based on MCAR and MAR estimates.

Robustness checks and identification assumptions

Once we obtain the comparison sample in Autumn 2019, we will look at the primary outcome during the pre-intervention period (2015-2017) per study group. We will conduct t-tests to assess whether the difference in the PSC raw scores between the treatment and comparison groups are statistically significant for each pre-intervention year. We will also produce a graphical analysis of pre-treatment trends in primary outcome between the groups.

Effect size calculation

Using the statistical analysis guidance provided by EEF, the formula for calculating effect size is given by:

$$\mathbf{ES} = \frac{(\bar{\mathbf{Y}}_{\mathbf{T}} - \bar{\mathbf{Y}}_{\mathbf{C}})\mathbf{adjusted}}{\mathbf{s}^{*}}$$

The numerator for the effect size calculation will be the coefficient of the treatment group from the multi-level model. All effect sizes will be calculated using total variance from a multilevel model, without covariates, as the denominator i.e. equivalent to Hedges' g.

Confidence intervals for each effect size will be derived by multiplying the standard error of the interaction coefficient by 1.96. These will be converted to effect size confidence intervals using the same formula as the effect size itself.

Implementation and process evaluation

The implementation and process evaluation (IPE) will complement the impact evaluation and gather information of importance to understand the implications for further application. It will cover all eight dimensions and five implementation factors set out in the introductory handbook (Humphrey et al., 2016), addressing the research questions below.

To minimise the burden on schools and avoid unnecessary expenditure, we propose to combine methods and instruments where possible, to explore the IPE (for example, to use Management Information (MI) data required for TLIF projects to monitor participation). The process evaluation will focus on wave 2 schools, although we will collect some retrospective end-point data from wave 1 schools (fidelity MI from the developer and teacher/TA survey).

Research Questions

RQ1 Is fidelity to the intervention maintained? What was delivered, extent of adherence to treatment approach and what the intervention replaced in treatment schools

RQ2 How much does dosage differ across the sample? I.e. participation in training and development days by relevant staff; frequency and duration of phonic lessons delivered in Reception and Year 1

RQ3 To what extent do participants engage with the intervention? Response of staff and pupils to the intervention, implementation challenges and adaptations and the reasons for these.

RQ4: To what extent is the intervention distinctive from existing literacy practice? Business as usual versus intervention group practice.

RQ5: What level and type of support does RM provide to intervention schools? Use and quality of support provided by the developers, variation between schools, impact on engagement and 'success'

IDEA workshop

While covering the usual aspects of the Tidier framework and understanding the Theory of Change for RWI, the IDEA workshop, held in February 2018, gave the opportunity to confirm plans for CPD observation/provider interviews and refine the methods of data collection for the IPE to correlate with the developer's already-established activity. This included the fidelity questionnaire, which was reviewed during the IDEA workshop to ensure that it is fit for purpose as a fidelity measurement tool for the evaluation. We also considered timing of this and other measures to avoid overburdening participating teachers, and how to sample for the other measurement tools to reflect the range of participating schools. The IDEA workshop was wide-ranging and exploratory in nature, and as such took all fidelity dimensions into consideration.

Fidelity questionnaire

The fidelity questionnaire is a pre-existing tool used by RM as a method of measuring engagement and adherence with/to the intervention.

As an established tool that succeeds in informing the developers' view of adherence and engagement in order to intervene, it will also function well as a measurement tool of fidelity for the evaluation. As part of the intervention outside of the evaluation, it is updated by consultant trainers after each training session/development day; it is intended that this delivery method will continue into the evaluation. It covers the following dimensions;

- Fidelity: attendance at training and development days, number of hours for RL role, consistency of delivery and duration of one-to-ones
- Reach: which staff receive training and deliver the intervention

Teacher Surveys

Treatment Group Baseline survey

Staff delivering the intervention in cohorts 3 and 4 will complete a baseline survey in September 2018. This survey will be delivered online (along with an Excel spreadsheet version) to the Reading Leader, who will disseminate to staff who are delivering the intervention in school. It will cover the following areas:

- Teachers' existing knowledge and confidence in teaching phonics, reading and writing
- The school's existing strategies for supporting children who are falling behind (including 1:1 tuition)
- Any other literacy programme or CPD the school has been involved with

The survey will also use the core questions from the Teaching and Leadership Innovation Fund's (TLIF) evaluations. These include questions on:

- Quality of leadership and teaching
- Schools' approach to professional development
- School culture and staff satisfaction.

Treatment Group Endpoint survey

All staff delivering the intervention, across all four cohorts, will complete an end-point survey in May 2020. The survey will include the questions asked at baseline, to measure change in knowledge and confidence after the initiative, and those areas targeted by the TLIF core questions. Additional questions will be included at end-point, exploring the following dimensions:

- Fidelity and dosage; length/frequency of phonics lessons and one-to-one delivery

- Quality; level of preparedness to deliver within the school following training, and the amount of support given/received within the school
- Reach: which staff receive training and deliver the intervention, any interference in the teaching group (e.g. regular music lessons that interrupt teaching for certain pupils)
- Programme differentiation: how distinct the RWI programme is from any existing practice (or business as usual), what other literacy support pupils receive
- Adaptation: changes, if any, made to the intervention during implementation and why
- Responsiveness: do teachers/teaching assistants enjoy delivering the intervention; perceptions of pupil engagement, confidence, motivation
- Programme support: quality of support given by the developer, including the portal, phone and email support and training
- Costs: direct marginal costs of intervention and staff time

Middle and senior leaders will be asked to complete a small number of additional questions asking for their views of the impact of RWI at school level.

Comparison group survey

There will be an endpoint proforma for targeted teachers in comparison schools in May 2020. As these schools will only receive this survey, this is intended to gather information about the school's literacy provision and business as usual throughout the duration of the intervention period (the preceding 2 academic years). Targeted dimensions are:

- Programme differentiation: teachers identify normal/existing practice, what literacy support pupils receive and the number of hours of literacy teaching per week (including one-to-one)
- Monitoring comparison; identify any change that has occurred over the intervention period and its drivers. Identify any unusual or unique characteristics of the comparison cohort
- Costs: direct marginal costs of Business as Usual including any other programmes the school has taken part in during the intervention period, taking into account the higher costs of the TLIF schools associated with the sample.

Observations and Interviews

Training observation

The evaluation team will observe the complete set of CPD delivery received by one school, ensuring a complete picture of the scope of training. The observations will ensure the team fully understand the intervention materials and expectations of teachers/schools. Targeted dimensions are:

- Fidelity: the extent to which trainers adhere to the set training materials and timings, or shape training to fit the audience (or for other reasons)

- Quality and responsiveness: the quality of training provided, including the level of attendee engagement, the clarity and overall strength of the trainer

Developer interviews

We will conduct telephone interviews with five developers/trainers, including the owner of the programme/company Ruth Miskin Training; the two members of the development team that have been involved in planning, designing and delivering the intervention; and two further trainers, identified by the developers. Interviews will further explore the intervention characteristics including expected dosage, CPD and support arrangements, expected classroom implementation and permissible tailoring, and any planned changes to implementation. Interviews will take place in September-October 2018 and will mainly focus on the following dimensions:

- Adaptation: where and why it may be necessary to adapt the training sessions and how this may affect fidelity
- Programme support: how the trainer/developer identifies who needs more support; methods of receiving requests for help; how much this has been a feature of the intervention in previous years
- Cost: the cost of the programme, both from a school's perspective and from the developer's

Staff interviews

We will conduct telephone interviews with a total of 30 school staff across 10 schools in Spring 2020, to include one senior leader, one teacher and one teaching assistant. Interviews will cover all fidelity dimensions, with a particular focus on the following;

- Preplanning and foundations: What was the existing level of need, readiness and capacity for teaching phonics and reading in recruited schools?
- Implementation support: What training and support was provided and how was this perceived?
- Implementation environment: How does RWI differ from schools' usual practice; senior leader support and any barriers to delivery?
- Implementer factors and adaptation: who delivers the interventions (e.g. teachers, TAs) and what experience/training have they had; what adaptations have staff made and why?

Cost evaluation

The cost of programme delivery will be explored from the school's and developer's perspectives. Information will be collected about the cost of the intervention as it was

delivered in the evaluation, and about what it would cost a school to self-fund the entire costs of receiving and delivering RWI. As the programme is partially funded for intervention schools by the TLIF, further cost information will be sought from the DfE if needed. Costs will then be calculated as a cost per pupil from the school's perspective, as if schools were paying for the intervention, based on marginal financial costs. This will reflect the fact that the cost for schools will be higher in this project due to the nature of the sample defined by TLIF. We also propose to collect Business as Usual (BaU) cost data from the comparison group, including any other phonics or literacy interventions the school may have received during the intervention period.

Questions will be administered in the surveys and during the telephone interviews mentioned above, and during the telephone interviews with the development team. We will explore direct, marginal costs including: training costs, staff salary costs if over and above the hours of current staff; purchasing costs for resources, meals, subsistence, travel and any out of hours room hire.

We will also report 'time' in terms of the amount of hours spent by staff and any other volunteers in preparing and delivering the intervention; and any re-allocation of existing resources (e.g. allocation of a named contact for the programme). We will report pre-requisite costs, which may include reading or phonics resources which a school may already have. RWI will be considered within the wider context of the costs of other literacy support programmes; taking into account existing costing methods and published costs (Curtis, 2013).

Ethics

The study will be designed, conducted and reported to CONSORT standards (http://www.consort-statement.org/consort.statement/) and registered on http://www.controlled-trials.com/. The evaluation will be conducted in accordance with NFER's Code of Practice, available at: http://www.nfer.ac.uk/nfer/about-nfer/code-of-practice/nfercop.pdf. NFER, Ruth Miskin Training and EEF will work together to ensure each organisations' policies can be applied in practice.

Ethical agreement

Ethical agreement for participation within the evaluation will be provided by the headteacher of the school. Parents will be provided with full details about the intervention, and will be given the opportunity to withdraw their child from data processing if they have objections to this.

All data gathered during the trial will be held in accordance with the Data Protection Act (1998), and from May 2018 with the General Data Protection Regulation (2018), and will be treated in the strictest confidence by the NFER, EEF, and Ruth Miskin Training. Pupil data collected from schools by NFER will not be made available to anyone outside of those parties listed. Our legal basis for gathering and using this data is legitimate interest, through our work as a research organisation.

Data protection

The legal basis for processing the personal data accessed and generated by the evaluation is covered by GDPR Article 6 (1) (f) which states that;

> 'processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party except where such interest are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of the personal data'.

We have carried out a legitimate interest assessment which demonstrates that the evaluation fulfils one of NFER's core purposes (undertaking research, evaluation and information activities) and is therefore in our legitimate interest, that processing personal information is necessary for the administration of the quasi-experimental design evaluation. We have considered and balanced any potential impact on the data subjects' rights and find that our activities will not do the data subjects any unwarranted harm.

In setting out the roles and responsibilities for this trial, the three parties (NFER, Ruth Miskin Training and the EEF) have signed a Data Sharing Agreement. This includes a description of the nature of the data being collected and how it will be shared, stored, protected and reported by each party. In addition, NFER will provide a memorandum of understanding to schools, explaining the nature of the data being requested of schools and children, how it will be collected, and how it will be passed to and shared with NFER.

For the purpose of research, UPN and test outcome data for all pupils in the trial will be linked with information about pupils from the National Pupil Database (held by the DfE) and other official records, and shared with NFER, DfE, EEF, Ruth Miskin Training, EEF's data archive contractor FFT Education, and, potentially, in an anonymised form to the UK Data Archive. Pupil data will be treated with the strictest confidence. Neither we, nor any of the named parties, will use pupil names or the name of any school in any report arising from the research.

On conclusion of the project, the Fischer Family Trust (see http://www.fft.org.uk/) will collate and de-identify the data for upload to the EEF data archive. The archived data will be available in a de-identified form with restricted access for research purposes only. NFER handles personal data in accordance with the rights given to individuals under data protection legislation. Individual rights are respected.

For further information, please see the Privacy Notice for the Evaluation of Ruth Miskin Training: *Read Write Inc*, available at https://www.nfer.ac.uk/media/3018/efrw_privacy_notice.pdf

Personnel

Name	Institute	Roles and responsibilities

Simon Rutt (SR)	NFER	Project Director, responsible for leading the NFER team and project delivery.
Gemma Stone (GS)	NFER	Project manager, responsible for overseeing the day to day running of the trial
Sarah Lynch	NFER	Process evaluation lead, responsible for managing the process evaluation activities and analysis
Caroline Sharp	NFER	Process evaluation director, responsible for overseeing the development of IPE tools
Kathryn Hurd (KH)	NFER	Test and Schools administration lead, responsible for overseeing recruitment, school contact and testing
Afrah Dirie (AD)	NFER	Statistician, responsible for statistical analysis

Risks

Risk	Likelihood/ impact	Mitigation
Insufficient 'similar' schools available to be recruited to the comparison group	High/High	Eligibility criteria will be expanded in close consultation with EEF and DfE where necessary.
Contamination within comparison group	Medium/High	By recruiting late within the trial we can exclude any schools already receiving RWI. Surveys will ask teachers to identify if they are already using any Ruth Miskin Training materials, at which point teachers can be omitted from analysis. A 10% over-recruitment will ensure any withdrawals do not damage power.
Lack of response from schools at post/follow-up	Medium/High	Engagement, contact and support from delivery team will be maintained after the intervention and until follow-up completed. MoU will clearly outline responsibilities of schools.
Researcher loss	Medium/Medium	NFER has a large research department with numerous researchers experienced in evaluation who could be redeployed.
Incomplete data returned by schools	Medium/Medium	MoU sets out clearly what is expected in terms of data collection at each time point. NFER will use reminding strategies to support schools to provide data. Developer will support NFER with encouraging schools to complete and return data.

Timeline

Date	Activity	Staff responsible/ leading
Feb – July 2018	• Setup meetings, IDEA workshop, TIDieR and Theory of Change developer • Study plan developed • Contracts and agreements setup	Simon Rutt, Gemma Stone Afrah Dirie
July – Aug 2018	• Development of IPE instruments • Development of school engagement materials	Sarah Lynch Caroline Sharp
Sep 2018	• Treatment schools in all cohorts complete and return Memorandum of Understanding • Cohort 2, 3 and 4 treatment schools begin intervention • Cohort 3 and 4 treatment schools undertake baseline survey	Kathryn Hurd Gemma Stone Sarah Lynch
Oct – Dec 2018	• Training continues with observation • Telephone interviews with developers • Fidelity questionnaire completed and regularly updated by staff in all cohorts	Gemma Stone Sarah Lynch
Jan – May 2019	• Training continues with observation	Sarah Lynch
Sep 2019	Year 2 of programme commences	
Sep – Dec 2019	• Preparation of comparison group sample • Preparation of comparison group school engagement materials	Simon Rutt Afrah Dirie Kathryn Hurd Gemma Stone
Jan – Feb 2020	• Recruitment of comparison group • Pupil level data collected across all treatment and control schools	Kathryn Hurd
Mar 2020	NPD request	Gemma Stone
Mar-May 2020	Telephone interviews with school staff	Sarah Lynch
May – June 2020	• Comparison schools undertake comparison (BAU) survey • Cohorts 1 and 4 sit Key Stage 1 assessments • Cohort 2 and 3 sit Phonics Screening Check	Gemma Stone Kathryn Hurd Sarah Lynch
June 2020	Collection of raw scores (KS1 and PSC) from all schools	Kathryn Hurd
June – July 2020	Treatment group (all cohorts) undertake end point teacher survey	Gemma Stone Kathryn Hurd Sarah Lynch
August 2020	Data send to NPD for matching	Afrah Dirie
Sep – Nov 2020	Impact analysis Process analysis	All
Nov 2020	Emerging findings meeting	All

Nov 20 – Jan 21	Report writing	All
Jan 31 2021	Draft report delivered	Gemma Stone
Mar 2021	Comments on draft report and amendments made	Gemma Stone

References

Abadie, Alberto and Javier Gardeazabal (2003). The economic costs of conflict. A case study of the Basque Country. The American Economic Review 93(1): 113-132.

Abadie, Alberto; Alexis Diamond; and, Jens Hainmueller (2010). Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. Journal of the American Statistical Association 105(490): 493-505.

Abadie, Alberto; Alexis Diamond; and, Jens Hainmueller (2015). Comparative politics and the synthetic control method. American Journal of Political Science 59(2): 495-510.

Angrist,J.D., & Imbens, G.W. (1995) Two stage-least squares estimation of average causal effects in models with variable treatment intensity. Journal of the American Statistical Association 90,431-442.

Austin, Peter (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research, 46(3): 399-424,

Austin, Peter (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. Pharmaceutical Statistics 10: 150-161

Austin, Peter (2014). A comparison of 12 algorithms for matching on the propensity score. Statistics in Medicine 33: 1057-1069.

Austin, Peter; P. Grootendorst and G. Anderson (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. Statistics in Medicine 26(4):734-753.

Baser, Onur (2006). Too much ado about propensity score models? Comparing methods of propensity score matching. Value in Health 9(6): 2006

Bates, Douglas; Maechler, M.; Bolker, B.; and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi:10.18637/jss.v067.i01

Beaujean, A. Alexander (2012). BaylorEdPsych: R Package for Baylor University Educational Psychology Quantitative Courses. R package version 0.5. URL: https://CRAN.R-project.org/package=BaylorEdPsych

Buuren, Stef van and Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. URL: https://www.jstatsoft.org/v45/i03/.

Caliendo, Marco and Sabine Kopeinig (2005). Some practical guidance for the implementation of propensity score matching. DIW Discussion Papers No. 485

Carpenter, J. R., and Kenward, M.G. (2007) Missing Data in Randomised Controlled Trials: A practical guide http://missingdata.lshtm.ac.uk/downloads/rm04_jh17_mk.pdf

Department for Education (2013) National Curriculum for England. [online] Available: https://www.gov.uk/government/publications/national-curriculum-in-england-english-programmes-of-study

DfES (2010) The importance of teaching schools: white paper 2010 (online). Available: https://www.gov.uk/government/publications/the-importance-of-teaching-the-schools-white-paper-2010

Education Endowment Foundation (2015). *Improving Numeracy and Literacy.* London: EEF [online]. Available: https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Campaigns/Evaluation_Reports/EEF_Project_Report_ImprovingNumeracyAndLiteracyInKeyStage1.pdf
[28 July 2018]

Education Endowment Foundation (2018). *Statistical analysis guidance for EEF evaluations*. London: EEF [online] Available: https://educationendowmentfoundation.org.uk/public/files/Grantee_guide_and_EEF_policies/Evaluation/Writing_a_Protocol_or_SAP/EEF_statistical_analysis_guidance_2018.pdf
[12 July 2018]

Ho, Daniel. E; K .Imai; G. King; and E.A. Stuart (2013). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. Journal of Statistical Software, Vol. 42, No. 8, pp. 1-28. URL http://www.jstatsoft.org/v42/i08/

Jiang, Yang and Small, Dylan (2014). ivpack: Instrumental Variable Estimation. R package v ersion 1.2. https://CRAN.R-project.org/package=ivpack

King, Gary; C. Lucas; and, R. Nielsen (2017). The Balance-Sample Size Frontier in Matching Methods for Causal Inference. American Journal of Political Science 61(2):473-489.

Lee, David and Thomas Lemieux (2010). Regression Discontinuity Designs in Economics. Journal of Economic Literature 48: 281-355.

Linde, Ariel and Paul Yarnold (2016). Combining machine learning and matching techniques to improve causal inference in program evaluation. Journal of Evaluation in Clinical Practice 22(6): 868-874.

Little, Todd.D (2014) The Oxford Handbook of Quantitative Methods in Psychology. Oxford University Press.

Liu, Chengfang, L.Zhang, R.Luo, S.Rozelle, and P.Loyalka. (2010). The effect of primary school mergers on academic performance of students in rural China, International Journal of Educational Development (30): 570–585

McClelland, Robert and Sarah Gault (2017). The synthetic control method as a tool to understand state policy. The Urban Institute research report.

McCrary, Justin (2008). Manipulation of the running variable in the regression discontinuity design: A density test. Journal of Econometrics 142(2): 698-714.

McKnight, P.E., McKnight, K.M., Sidani, S., and Figueredo, A. J. (2007) Missing Data: A gentle introduction. The Guildford Press

Ofsted (2010) *Reading by Six: How the Best Schools Do It.* London: Ofsted. *[online]* Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/379093/Reading_20by_20six.pdf

Rose, J. (2006) *Independent Report on the Teaching of Early Reading: Final Report,* London: DfES.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.

Stuart, Elizabeth (2010). Matching methods for causal inference: A review and a look forward. Stat Sci 25(1):1-21.

Torgerson, C.J., Brooks, G. and Hall, J. (2006) *A Systematic Review of the Research Literature on the Use of Phonics in the Teaching of Reading and Spelling,* DfES Research Report 711, London: DfES.