Flexible Phonics
Statistical Analysis Plan

Evaluator (institution): Institute for Employment Studies
Principal investigator(s): Anneka Dawson and Helen Gray

Education Endowment Foundation

Template last updated: August 2019

PROJECT TITLE	Flexible phonics, a two-armed cluster randomised trial
DEVELOPER (INSTITUTION)	University College London (UCL), Institute of Education (IoE)
EVALUATOR (INSTITUTION)	Institute for Employment Studies (IES)
PRINCIPAL INVESTIGATOR(S)	Anneka Dawson and Helen Gray
SAP AUTHOR(S)	Anneka Dawson, Helen Gray and Clare Huxley
TRIAL DESIGN	Two-arm cluster randomised controlled trial with random allocation at school level
TRIAL TYPE	Efficacy
PUPIL AGE RANGE AND KEY STAGE	Reception year age 4-5 (Early Years)
NUMBER OF SCHOOLS	123
NUMBER OF PUPILS	2,706
PRIMARY OUTCOME MEASURE AND SOURCE	York Assessment for Reading Comprehension Test - word recognition subscale
SECONDARY OUTCOME MEASURE AND SOURCE	York Assessment for Reading Comprehension Test full score, Mispronunciation Communication Test, delayed post- test of Phonics screening test at end of Year 1

SAP version history

VERSION	DATE	REASON FOR REVISION
1.2 [*latest*]		
1.1		
1.0 [*original*]		*N/A*

Table of contents

Flexible Phonics
Statistical Analysis Plan

Evaluator (institution): Institute for Employment Studies
Principal investigator(s): Anneka Dawson and Helen Gray

Template last updated: August 2019

Introduction

THE INTERVENTION

The Flexible Phonics intervention seeks to give Reception class teachers and Teaching Assistants new strategies to help all children learn to read. It is designed to complement existing phonics programmes, allowing them to be delivered broadly as usual. The Flexible Phonics intervention gives children more strategies to flexibly read all words and is expected to be particularly powerful in enabling children to independently read novel exception words (words that break phonic rules such as 'the', 'two', 'between', 'above', etc.). Children will learn how to use phonics in close conjunction with authentic children's texts to become confident, motivated, readers.

Teachers and Teaching Assistants will be trained in two main strategies:

1. Direct Mapping (DM). Children will be taught grapheme to phoneme correspondences (GPCs) and then given texts to read that include examples of the GPCs that they have just learned. Initially, these will be carefully selected pre-existing decodable texts, or specifically crafted controlled texts, but real books will be introduced slowly and strategically. While many models of phonics teaching link phonics and texts, DM aims to do so more thoroughly, consistently, and on the same day as children learn the specific GPCs, aiming to ensure that children understand phonics in context.
2. The second strategy, Set-for-Variability (SfV), teaches pupils to add in another step after they have blended phonemes to graphemes where pupils 'set-for-variability'. Pupils are encouraged to consider what the word may be, given both the distance between these blended sounds and known words, and potential spelling to sound inconsistencies. This enables children to better recognise all words but can be especially useful when learning to recognise exception words (e.g. 'wasp').

These techniques will be taught to pupils as part of normal phonics lessons.

Schools allocated to the treatment group will receive free children's books to the value of £400 per school which can be used to implement the strategies. Teachers and Teaching Assistants will receive three half days of training delivered remotely, a copy of a Teacher Manual and associated resources and three follow-up on-line appointments. Appointments will be offered to the Reception teacher/ TAs teaching Flexible Phonics in the school. Each class within the year group will be offered a separate appointment lasting around 30-minutes or a group appointment if they prefer. Staff who teach the classes selected for pre and post tests have been a particular focus for follow-up support. Appointments will enable staff to ask questions and get advice on best practice implementation of the programme.

As well as the initial training and follow-up sessions, there will also be an online platform with resources including videos of the training sessions, short videos of key lessons, audio files for some of the teaching activities, the training manual, FAQs, slides and any other training

documents and resources shared by schools. This platform will also include a discussion board for all trained teachers and TAs to join and they can ask for 'as and when' additional support through the discussion board if needed. Best practice and resources provided by partner schools will be shared on schools' behalf by the Flexible Phonics Team through this medium.

A monthly newsletter will be sent to schools to showcase any new resources contributed by partner schools (these can be accessed directly from the newsletter), share good practice, highlight any relevant articles on topics of concern for schools and to share answers to frequently answered questions raised during the training and in online appointments more widely.

Proactive support for schools will be provided by the Flexible Phonics Support Team by email between February and July between online appointments, where relevant resources and best practice will be shared proactively with the schools. Schools can also contact the Flexible Phonics Support Team by phone or email as needed.

Teachers can also choose to share videos of their own practice for feedback through video calls with UCL staff if they choose for specific further feedback.

THE TRIAL

The primary research question to be answered by the trial is:

- RQ1. Does the Flexible Phonics intervention improve Reception children's word reading ability? (measured by the York Assessment for Reading Comprehension (YARC) Early Word Recognition subscale)

The secondary research questions are:

- RQ2. Does the Flexible Phonics intervention improve Reception children's literacy outcomes? (measured by more general literacy tests)
- RQ3. What is the differential impact of direct mapping and set- for-variability skills on children's word reading ability?
- RQ5. Does the Flexible Phonics intervention improve word reading ability differentially for children eligible for Free School Meals (FSM)?
- RQ6. Does the Flexible Phonics intervention improve word reading ability differentially for children of low ability?
- RQ7. Does the Flexible Phonics intervention improve Reception children's phonics skills one year later at the end of Year 1?

Initially the aim was to answer a further secondary research question on whether the Flexible Phonics intervention provides value- added improvement to Reception children's word reading ability compared to good phonics teaching alone in schools identified with good phonics practice (RQ4)? However, it will not be possible to answer this research question as Year 1 Phonics Screening data was not collected from schools during recruitment and it is considered too much of a burden to collect this data from schools during the intervention in the context of the ongoing pandemic.

123 schools in the Greater London area agreed to take part in the trial, This area was specified by the funder EEF and was partly chosen for ease of delivery when the intention was to use face-to-face training. Schools have been randomly assigned to the treatment and

control groups. A random number was generated for each school and schools sorted on this number in ascending order. Each school was numbered from 1 to 123 based on this sort order and those with an odd number were assigned to the treatment group, and those with an even number assigned to the control group.

In schools with multi-form entry, one class was selected at random to take part in pre-intervention testing. This random selection of classes took place on 9 October 2020. A pre-test was then conducted and following completion of the pre-test schools were allocated to the treatment and control groups. This was done in two batches, on 3 and 9 December 2020, to allow UCL to share the information with the selected schools to enable them to book on to the early January training sessions before the Christmas half-term break. The approach to randomisation ensured that equal numbers of schools were assigned to the treatment and control groups in each batch (albeit with one additional school assigned to the treatment group overall, due to the fact that an odd number of schools participated in the trial). The section on 'Imbalance at baseline' explains the steps we will take to identify whether there are in fact any problems with the balance between the two groups. The three-part training sessions (each one of 3 hours duration) were held between January and February and the intervention will be delivered until the end of July 2021. Most schools started teaching Flexible Phonics once classroom teaching resumed in March 2021. From this point until the end of the Summer term all phonics lessons (normally three to four a week, depending on the school) will incorporate Flexible Phonics strategies. Post-intervention testing will take place in June and July 2021.[1]

The analysis will test the efficacy of the flexible phonics intervention. The primary analysis will focus on estimating the intention-to-treat effect. This will be supplemented by a compliance analysis to estimate the impact of Flexible Phonics in schools observed to be fully compliant with the intervention. Subgroup analyses will also be used to estimate the differential impact of Flexible Phonics on pupils eligible for Free School Meals (FSM) and separately on low-ability pupils. Subgroup analysis will also be used to explore the impact of the intervention in schools with existing good phonics practice.

We will explore the impact of the two strategies of Direct Mapping and Set for Variability using causal path analysis. The MCT measure is expected to capture the direct impact of Set for Variability, whilst the primary outcome measure of Early Word Recognition is expected to be affected by both Set for Variability and Direct Mapping. The path analysis will seek to disentangle the relative contribution of each strategy..

[1] The implications of school closures between January and March 2021 due to the pandemic were discussed with the delivery team, but it was decided that while school closure clearly affects performance, for various reasons this should not undermine or fatally compromise the ability to observe any impact from Flexible Phonics by the time of the post-test.

Design overview

Trial design, including number of arms		Two-arm, cluster randomised control efficacy trial with pupil-level outcomes
Unit of randomisation		School
Stratification variables (if applicable)		None
Primary outcome	variable	Early Word Recognition
	measure (instrument, scale, source)	Early Word Recognition subscale raw score (0-30) from the York Assessment for Reading Comprehension (YARC)
Secondary outcome(s)	variable(s)	Early Word Reading composite measure Mispronunciation Correction Literacy over the longer-term
	measure(s) (instrument, scale, source)	For literacy: - The sum of standardised scores derived from each of the four YARC subscales i.e. early word recognition, letter sound knowledge, sound deletion and sound isolation. - Score on the Year 1 Phonics Screening check for longer-term outcomes. For Mispronunciation Correction: An adapted version of Tunmer and Chapman's Mispronunciation Correction Test (2012) as used in Dyson et al. (2017) using the words most commonly used in English children's books.
Baseline for primary outcome	**variable**	Early Word Recognition
	measure (instrument, scale, source)	Early Word Recognition subscale raw score from YARC
Baseline for secondary outcome	**variable**	Early Word Recognition and Letter Sound Knowledge composite measure
	measure (instrument, scale, source)	Constructed from the standardised scores for the Early Word Recognition and Letter Sound Knowledge subscales from YARC

Sample size calculations overview

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
Minimum Detectable Effect Size (MDES)		0.23 Standard deviations	0.37 Standard deviations	0.21 Standard deviations	0.33 Standard deviations
Pre-test/ post-test correlations	level 1 (pupil)				
	level 2 (class)	0.4	0.4	0.4	0.4
	level 3 (school)				
Intracluster correlations (ICCs)	level 2 (class)				
	level 3 (school)	0.15	0.15	0.15	0.15
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		Two-sided	Two-sided	Two-sided	Two-sided
Average cluster size		23[2]	3	22	3
Number of schools	intervention	50	50	62	62
	control	50	50	61	61
	total	100	100	123	123
Number of pupils	intervention	1,150	150	1,364	186
	control	1,150	150	1,342	183
	total	2,300	300	2,706	369

The sample size calculations at protocol stage were based on an assumption that an average of around 23 pupils from each participating school would take part in testing during both the pre- and post-test phase. As mentioned in the previous section, one class was selected at random for testing. The sample size calculations allowed for attrition or withdrawal by around 15 per cent of pupils from an infant class of average size. This is similar to the rate of attrition seen in the evaluation of Abracadabra – an online literacy programme for Year 1 pupils (McNally, Ruiz-Valenzuela and Rolfe, 2016).[3]

Around 15 per cent of nursery and primary school children in Inner and Outer London were known to be eligible for and claiming free school meals in January 2018. Allowing for attrition between pre- and post-testing, this equates to around three children per class. In practice, early analysis of data on the 123 schools participating in the Flexible Phonics trial shows that they have an average of 26 pupils. With a 15 per cent attrition rate, this would reduce the achieved sample to 22 pupils per class, although the likely numbers claiming free school meals would remain at around 3 pupils per class.

[2] This is based on the expected achieved sample size i.e. after attrition.
[3] DfE figures indicated that the average infant class contained 27 pupils in 2018 (DfE 2018: 11).

The calculation of the MDES assumes that the pre-test explains 16 per cent of the variation in the post-test scores. This is based on evidence from the Nuffield Early Language Intervention project which found a correlation between pre- and post-test scores for our primary outcome measure, the early word recognition subscale from the York Assessment of Reading for Comprehension, of 0.39.

The MDES calculation also assumes that the intra-class correlation is 0.15. Previous EEF evaluations on early years have indicated that schools explain around 15 per cent of the variation in pupil attainment. For example, the efficacy trial of EasyPeasy found an ICC of 0.18 (Robinson Smith *et al.*, 2019), whilst the efficacy trial of Family Skills reported an ICC of 0.15 at class level (Husain *et al.*, 2018). The sample size calculations are based on the standard assumptions of 80% power and 5% significance level and an even split in the number of schools in the treatment and control groups.

When the protocol was drafted, the expectation was that around 100 schools would participate in the trial. This resulted in an MDES on the primary outcome measure of 0.23 standard deviations and 0.37 standard deviations for the subset of pupils eligible for free school meals. In practice, 123 schools were recruited to the trial and were randomised to the treatment and control groups. Of these, 62 schools were randomised to the treatment group and 61 to the control group. This reduces the MDES on the primary outcome measure to 0.21 standard deviations, or 0.33 standard deviations for the free school meals sample. The size of the MDE for the free school meals subgroup means that it is unlikely to be possible to discern whether the intervention has had a clear impact on this subset of pupils.

Analysis

The main analysis will compare the primary and secondary outcome measures for all schools assigned to either the treatment or control groups to estimate the impact of the intention to treat. The primary outcome measure will be the raw score on the Early Word Recognition subscale from the YARC. The secondary outcome measures are as follows:

- the sum of the standardised scores derived from the four YARC subscales (early word recognition, Letter Sound Knowledge, Sound Deletion and Sound Isolation),
- the raw score on an adapted version of Tunmer and Chapman's Mispronunciation Correction Test (2012) as used in Dyson et al. (2017) based on the words most commonly used in English children's books and
- the raw score on the Year 1 Phonics Screening check carried out one year after the post-intervention test, to capture longer-term outcomes.

The first of the secondary outcome measures will be constructed from standardised scores on each of the subscales to allow for the fact that each contains a different number of items. The other two secondary outcome measures will be based on raw scores as each consists of a single scale.

The analysis will control for the prior attainment of individual pupils to increase statistical power and the precision of the impact estimate (following EEF guidance). Outcomes will be measured at individual pupil level, but the analysis will use multi-level modelling to take into account the nested structure of the trial, with pupils clustered within schools, with the schools allocated either to the treatment or the control group.

The measure of prior attainment will be based on pre-test scores. For the primary outcome measure of Early Word recognition from the YARC the prior attainment measure will be the

raw pre-test score on the Early Word Recognition subscale. For the secondary outcome measures, a composite measure derived from the pre-test standardised scores for the YARC Early Word Recognition and the Letter Sound Knowledge subscales will be used as the measure of prior attainment. Again, the scores will be standardised before they are combined to reflect the fact that the two subscales have different numbers of items.

Estimated impacts will be converted into Hedges' *g* effect size (1981) which uses the estimated total pooled standard deviation of the treatment and control groups. This provides a more conservative estimate of impact compared with using the within-school pooled standard deviation. Hedge's *g* effect sizes will be reported along with 95 per cent confidence intervals, as per EEF reporting guidelines and the analysis will explore whether impact estimates are statistically significant at the 5 per cent level or better.

Primary outcome analysis

The primary outcome measure to be used to assess the impact of the Flexible Phonics intervention is the Early Word Recognition subscale raw score (0-30) from the York Assessment for Reading Comprehension (YARC). Children are asked to read 30 single words which are graded in terms of difficulty. The Early Word Recognition subscale is a measure of overall literacy and is thus thought to be the measure more likely to be affected by the Flexible Phonics programme given the age of the children at the time of pre- and post-intervention testing.

The option of using other YARC subscales individually or in combination with each other was considered. However, the individual subscales were expected to vary in terms of whether they were likely to be affected by the intervention, or suitable for Reception age children at both the pre- and post-test points. As the YARC was developed for children across the 4 to 7 age range, there was a risk of floor effects from a large proportion of children achieving low scores on some of the individual subscales, particularly at the pre-test point.[4] The Early Word Recognition subscale was the measure identified as being most likely to capture any differences in prior attainment at the pre-test point for Reception age children. This was partly because it includes a larger number of items than any of the other YARC subscales, and partly because it was thought most likely to be affected by Flexible Phonics techniques.

Children will be tested using the Early Word Recognition subscale both before and after participation in the Flexible Phonics programme so that this measure can be used to control for prior attainment when estimating the impact of the intervention on the primary outcome.

The analysis will use multilevel modelling to reflect the likelihood that there are similarities in the prior attainment and outcomes of pupils clustered within the same school. Failing to take into account between-school variance would be likely to mean that the standard errors around the impact estimates were understated, making it more likely that Flexible Phonics was judged to be effective when in practice the finding was due to the approach to estimation. A reanalysis of data from a number of EEF evaluations using different approaches to estimation suggested that Bayesian multi-level modelling with weakly informative priors is likely to improve the precision of the impact estimates compared with other techniques (Xiao, Kasim and Higgins, 2016). Given this finding, the analysis for the current study will be carried out using the crtbayes command available in the EEFanalytics.ado package developed for Stata 16. Stata's graphical diagnostics (including

[4] This issue was identified in the Tips by Text trial, although the report on the analysis has not yet been published.

trace, histogram, autocorrelation, and kernel density plots) will be used to check for MCMC convergence. The prior distribution will be based on the results of an earlier quasi-experimental study of Direct Mapping and Set for Variability in Canada (Savage *et al.*, 2018).

The equations to be estimated are as follows:

$$Y_{ijt} = \beta_0 + \beta_i Treat_i + \beta_2 Y_{ijt-1} + \eta_j + \varepsilon_{ij}$$

where i are pupils and j are schools (or more precisely the class randomised to the treatment or control group within each school), Y_{ijt} is the post-test Early Word Recognition score, Y_{ijt-1} is the pre-test Early Word Recognition score, $Treat_i$ is the treatment indicator (coded to 1 for individuals in the treatment group and 0 for individuals in the control group), η_j is a school-level random effect and ε_{ij} is the error term. Bayesian credibility intervals will be reported.

Secondary outcome analysis

As mentioned previously, three secondary outcomes will also be considered in the analysis:

1. The full test score from all four subscales within the York Assessment for Reading Comprehension. Results from each individual subscale will be standardised so that they each have a mean of zero and a standard deviation of one. Each of the standardised subscales will then be added together to derive a single measure of Early Word Reading.
2. The Mispronunciation Correction Test (MCT). This 40-item measure is suited to capturing the impact of the Set for Variability strategy. Fifteen pupils from the selected class in each school will be selected at random to take part in the MCT. Where the number of pupils eligible to be tested is 15 or fewer, all pupils will be asked to take the MCT. The analysis will estimate the impact of Flexible Phonics on the raw MCT score, which can range from zero to 40. The secondary outcome measure will also be used in a path analysis to seek to estimate the portion of any impact attributable to Direct Mapping, and the portion arising from the Set for Variability strategy.
3. The score on the Year 1 Phonics Screening check. This test is a statutory assessment administered to all Year 1 pupils. Pupils can attain a score between zero and 40. The cohort of Reception-aged children participating in the trial are expected to take the Phonics Screening check in June 2022, which is likely to be around 12 months after the end of participation in Flexible Phonics. Whilst the Phonics Screening check captures general literacy, rather than the specific skills that Flexible Phonics seeks to teach, the measure does provide a low-cost way of exploring whether any impact from Flexible Phonics on wider literacy is sustained. It is expected that the results of the Phonics Screening check will be available in the National Pupil Database in Autumn 2022. The analysis of the results of the Phonics Screening check is discussed in more detail in the section on the Longitudinal follow-up analyses.

As mentioned previously, for all of the secondary outcome measures the analysis would take into account prior attainment. Unlike the analysis of the primary outcome, this would be based on pre-test results on both the Early Word Reading and Letter Sound Knowledge subscales. The two subscales would be standardised and added together to produce a single measure of prior attainment.

In other respects, the approach to the analysis will be identical to the method used for the primary outcome, in that it will be based on Bayesian multi-level modelling with vague priors. Once again, Bayesian credibility intervals will be reported.

Subgroup analyses

The analysis will estimate the impact of Flexible Phonics on the following subgroups:

1. Pupils eligible for free school meals. This analysis will use the indicator of whether the pupil has been eligible for free school meals in the past six years from the National Pupil Database (EVERFSM_6_P).
2. Low-ability pupils. Low-ability pupils will be defined as those who score less than the median on the combined pre-test standardised Early Word Recognition and Letter-Sound Knowledge subscales.
3. Pupils in schools participating in the Nuffield Early Language Intervention (NELI) – as part of the government's COVID-19 support strategy. From the 62 schools allocated to the treatment group, 26 are participating in NELI, compared with 23 of the 61 schools assigned to the control group. A subgroup analysis will be used to determine whether the impact of Flexible Phonics varies depending on whether the school is also participating in NELI.

The subgroup analyses will involve interacting the subgroup identifier with the treatment group indicator to estimate the differential impact of Flexible Phonics on the primary outcome measure for pupils eligible for free school meals, for low ability pupils and for pupils at schools which are also participating in NELI. In addition, the impact of Flexible Phonics will be estimated for the subsamples of pupils eligible for free school meals, for low ability pupils and for pupils at schools not participating in NELI. In all cases the specification will be the same as that used in the main analysis of the primary outcome.

Additional analyses

A descriptive analysis will be used to explore whether the treatment and control groups differ across observable characteristics prior to the intervention. This is described in detail in the section on 'Imbalance at baseline'. If there are statistically significant differences at baseline, an alternative specification will be used to explore whether including these controls in the analysis affects the likelihood of detecting an impact from Flexible Phonics.

For 18 schools the pre-test was carried out remotely, and it is possible that this may also be necessary for some schools when conducting the post-test. The distribution of schools between the treatment and control groups where remote testing is necessary at pre- and/or post-tests will be assessed, as well as the characteristics of these schools compared with schools in either arm where face-to-face testing was possible. Having understood the potential impact of remote testing on the representativeness of the treatment and control groups, additional analyses of the primary outcome will be carried out to explore whether the main findings hold when excluding pupils who were tested remotely. In other respects, the approach to the analysis will be the same as that carried out for the primary outcome.

A path analysis will be carried out to seek to identify the contribution of the Direct Mapping and Set for Variability strategies to the overall effectiveness of Flexible Phonics (RQ 3). The MCT is expected to provide a direct measure of the impact of the Set for Variability strategy, whereas the Early Word Recognition subscale is thought to capture the impact of both strategies on overall literacy. A multilevel generalised path analysis will be used to calculate the portion of the impact of the intervention which can be attributed to the Direct Mapping strategy. The analysis will be done using the GSEM package in Stata which is appropriate

for the multi-level structure of the data. The analysis will be based on the assumption that the impact of flexible phonics as a whole on the primary outcome will be mediated by the impact of the Set for Variability Strategy on the score in the MCT. If the Set for Variability strategy has no impact on the MCT, it would also be assumed to have no impact on the score in the Early Word Recognition test. In this case, any impact on the primary outcome would be attributed solely to the Direct Mapping strategy.

Additional analyses will be used to explore the sensitivity of the impact estimates to the inclusion of pre-test scores. This will include repeating the main analyses on primary and secondary outcomes without pre-test scores and also imputing pre-test scores in cases where these are missing. The analyses of secondary outcomes will also be repeated using the pre-test Early Word Recognition raw score as the measure of prior attainment rather than the combination of both pre-tests. This will provide an insight into whether the estimated impact of Flexible Phonics on secondary outcomes is sensitive to the choice of pre-test measures.

Longitudinal follow-up analyses[5]

Pupils participating in the trial will take part in the national phonics screening check when they reach the end of Year 1. This will be around one year after the post-test which will be administered as part of the trial. This therefore provides an opportunity to assess whether Flexible Phonics has a lasting impact on literacy. It is expected that the results of the 2022 phonics screening check will be available by around September 2022.

The longitudinal analysis will be based on the intention-to-treat and will follow EEF guidance on longitudinal analysis pertaining when the analysis is carried out. If this differs from the approach to the analysis of the primary outcome measure, the longitudinal analysis will also include a secondary model which mirrors the analysis carried out for the primary outcome measure. The pupil's score on the phonics screening check will be used as the outcome measure.

Imbalance at baseline

As mentioned in the section on additional analysis, a descriptive analysis will be used to determine whether there are differences in the characteristics of the intervention and control groups prior to the intervention. The analysis will consider a range of pupil and school-level characteristics as follows:

- Pupil characteristics: Sex; eligible for FSMs; average age in months; pre-test scores.
- School characteristics: percentage of pupils eligible for FSMs; most recent Ofsted rating.

These pupil and school-level characteristics are expected to have a bearing on the likely impact of Flexible Phonics and so any imbalance in these characteristics between the treatment and control groups could potentially be expected to bias the impact estimates.

The assessment of imbalance will report absolute standardised differences between the two groups and between those falling into each of the different subgroups considered in the subgroup analysis. Any difference which are greater than 10 per cent will be highlighted as suggesting imbalance between the treatment and control groups on that particular

[5] Please see the longitudinal analysis guidance.

characteristic. The descriptive analysis will consider whether differences in the characteristics of the intervention and control groups are apparent at two time-points:

1. At the time of the pre-test. This will indicate whether the randomisation was successful in ensuring that the control group appeared similar to the control group prior to the intervention. It will therefore provide an insight into whether outcomes for the control group are likely to provide a credible estimate of the outcomes that the intervention group would have experienced if they had not participated in Flexible Phonics. Sizeable and statistically significant differences in the characteristics of the two groups prior to the intervention would reduce confidence that the impact estimates reflected the true impact of Flexible Phonics.

2. At the time of the final test. It is possible that pupils who do not take part in post-intervention testing experience different outcomes to those who remain in the study. The descriptive analysis will consider the balance between the intervention and control groups in the subset of pupils who take part in both pre- and post-tests.

The analysis will report means and standard deviations for continuous variables and counts and percentages in each category for categorial variables. Differences in pupil-level pre-test scores will be reported as effect sizes. The analysis will also show the correlation between pre- and post-test scores to show how these compare with the expectations set out in the sample size calculations.

Missing data

The analysis will provide details of the number of pupils participating in the trial for whom complete information is available across the list of pupil and school characteristics listed in the section on assessing Imbalance at baseline. This will be broken down by treatment arm. It will include an analysis of the numbers of pupils participating in both pre- and post-intervention testing and the numbers of missing observations at school and pupil-level. Binary variables will be derived to indicate missing observations on existing variables and probit regressions used to establish whether other variables predict the likelihood of a variable being missing. T-tests will also be used to identify statistically significant differences in the mean value of variables between individuals with missing and non-missing values on other variables.

The Stata mi suite of commands will be used to impute missing values jointly over clusters using the multivariate normal model. Whilst the main analysis will be based solely on full cases, additional analyses will be used to explore the sensitivity of the impact estimates to imputing missing data, whatever the scale of missings. Multiple imputation will be carried out using joint modelling to test the sensitivity of the main findings to imputing missing pre- and post-test data.

Compliance

A school will be considered to be compliant with the Flexible Phonics programme where:

A) The teacher of the class that has been selected for impact testing has attended all three training sessions (or watched the videos and attended a catch-up tutorial with Professor Savage or a Flexible Phonics Support Partner) and

B) Where teaching practice within that class is observed by a UCL Flexible Phonics Support Partner to have met the requirements in the rubric created by UCL (Global Treatment

Fidelity Rating, GTFR[6]). Originally, the Flexible Phonics Support Partners were going to observe teachers/TAs teaching phonics to their class during support visits to rate compliance but as support visits have moved to online, support assistants will rate compliance based on their discussions with schools during the three follow-up support sessions and support through email and other discussions, e.g. comments on the discussion boards. Schools have been invited to share videos of practice for feedback from the delivery team but this is optional and not required. It will only be possible to rate compliance in schools which participate in follow-up sessions and where the delivery team are able to assess teaching practice. This means that the compliance measure will not be available for schools which do not participate in these sessions, whether due to a lack of engagement, or confidence in delivering Flexible Phonics without the need for further input from the delivery team. This part of the compliance measure is examined on a 4-item scale, ranging from zero to 3: 0: No implementation of Flexible phonics, 1 Entry level: Some (but likely poor quality) implementation, 2 Adoption: Clear and competent regular delivery of intervention 3 Adaptive delivery: Expert and extended delivery of intervention. Some measures may not be relevant and so would be marked as 'Not applicable', e.g. a class of low performing readers may not be quite ready for print- based flexibility in mispronunciation correction of phoneme strings. The ratings for each measure will be the highest score seen over the course of the follow- up visits.

The compliance measure is based on an assessment of phonics practice through the follow-up sessions across five areas:

1. Direct Mapping. This involves linking grapheme to phoneme correspondences to texts which provide examples and using them across wider literacy and language activity. This is mandatory and must be given a score of 2 or 3 indicating they are delivering this in classes at an appropriate level [7].,
2. Oral flexibility – delivery of oral games to teach 'mispronunciation correction' of phoneme strings' (e.g. in games such as 'Simon says' or in wider classroom communications. This **or** print-based flexibility must also be given a score of 2 or 3 or Not applicable alongside Direct mapping.
3. Print-based flexibility in mispronunciation correction of phoneme strings (in print–based reading tasks and games and often linked to texts with high word frequency or in wider reading, this **or** oral flexibility must also be given a score of 2 or 3 or Not applicable alongside Direct mapping.

The attendance and outcome on the GTFR will be collapsed into a binary measure indicating that the school is either delivering the Flexible Phonics programme to the required standard, or cannot be considered to be compliant. Compliance will be defined at the level of the class. As noted above, it may not be possible for the delivery team to assess compliance in all cases if attendance at the support visits is limited, so it is likely that the compliance analysis will only be possible for a subset of schools in the intervention group. The compliance

[6] Please note that Vocabulary and continuous phonation also make up part of the GTFR for Flexible Phonics Support Partners to score but these do not factor into our compliance score and are used for UCL purposes only as they are not considered essential for compliance

[7] It is possible that by the time the intervention starts in the school year that all pupils have moved past the point direct mapping would be needed (they know all the GPCs) and if so, direct mapping could also be marked as not applicable, but this is regarded as very unlikely by the delivery team.

analysis will only be carried out if 50 or more schools in the treatment group can be assessed for compliance.

A Complier Average Causal Effect (CACE) analysis will be used to estimate the impact of Flexible Phonics on pupils who receive a version of the treatment which involves delivering the programme as intended (according to the measure described above) in the theory of change. This will use a two-stage least squares regression, with assignment to the treatment or control group used as an instrumental variable. In the first stage, the probability of compliance will be estimated (using the binary compliance measure). The second stage will use the predictions from the first stage to estimate the impact of compliance on the primary outcome measure of Early Word Recognition, controlling for pre-test scores in a similar way to the primary outcome analysis. The analysis will use the Stata command ivregress which can adjust standard errors to take account of the fact that pupils are clustered within schools.

Intra-cluster correlations (ICCs)

The ICC between the pre- and post-test data will be estimated at school-level. The correlation will be estimated using a hierarchical linear model without covariates and with school-level random effects using the following equation:

$$Y_{ij} = \beta_0 + \eta_j + \varepsilon_{ij}$$

where Y_{ij} is the pre- or post-test of individual *i* in school *j*, β_0 is a constant term, η_j is a school-level random effect and ε_{ij} is an individual-level idiosyncratic error term. The ICC will be estimated as follows:

$$ICC = \frac{var(\eta_j)}{var(\eta_j) + var(\varepsilon_{ij})}$$

Effect size calculation

Estimated impacts will be calculated in accordance with the EEF analysis guide to aid comparability with other trials. This involves estimating Hedges' *g* based on total variance, rather than within-cluster variance. This also gives a more conservative estimate of impact compared with using within-cluster variance. The effect size equation is as follows:

$$ES = \frac{\bar{Y}_t - \bar{Y}_c}{\sqrt{(\sigma_s^2 + \sigma_{error}^2)}}$$

Where ES is the estimated effect size, $\bar{Y}_t - \bar{Y}_c$ is the adjusted difference in mean outcomes between the treatment and control group and $\sqrt{\sigma_s^2 + \sigma_{error}^2}$ is the pooled unconditional variance of the treatment and control groups, taking into account both school level variance and pupil level variance.

Hedge's *g* effect sizes will be reported along with 95 per cent Bayesian credibility intervals, as per EEF reporting guidelines and the analysis will explore whether impact estimates are statistically significant at the 5 per cent level or better.

Bibliography

Dyson, H. *et al.* (2017) 'Training mispronunciation correction and word meanings improves children's ability to learn to read words.', *Scientific Studies of Reading*, 21(5), pp. 392–407. doi: 10.1080/10888438.2017.1315424.

Hedges, L. V. (1981) 'Distribution Theory for Glass's Estimator of Effect size and Related Estimators', *Journal of Educational Statistics*, 6(2), pp. 107–128. doi: https://doi.org/10.3102/10769986006002107.

Husain, F. *et al.* (2018) *Family Skills: Evaluation report and executive summary*. London: Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Family_Skills.pdf.

McNally, S., Ruiz-Valenzuela, J. and Rolfe, H. (2016) *ABRA: Online Reading Support Evaluation report and executive summary*. Education Endowment Foundation.

Robinson Smith, L. *et al.* (2019) *EasyPeasy: Learning through play. Evaluation report*. London: Education Endowment Foundation.

Savage, R. *et al.* (2018) 'Preventative Reading Interventions Teaching Direct Mapping of Graphemes in Texts and Set-for-Variability Aid At-Risk Learners', *Scientific Studies of Reading*, 22(3), pp. 225–247. doi: 10.1080/10888438.2018.1427753.

Tunmer, W. E. and Chapman, J. W. (2012) 'Does Set for Variability Mediate the Influence of Vocabulary Knowledge on the Development of Word Recognition Skills?', *Scientific Studies of Reading*, 16(2), pp. 122–140. doi: 10.1080/10888438.2010.542527.

Xiao, Z., Kasim, A. and Higgins, S. (2016) 'Same difference? Understanding variation in the estimation of effect sizes from educational trials', *International Journal of Educational Research*, 77, pp. 1–14. doi: 10.1016/j.ijer.2016.02.001.

York, B. N., Loeb, S. and Doss, C. (2018) 'One Step at a Time: The Effects of an Early Literacy Text Messaging Program for Parents of Preschoolers', *Journal of Human Resources*. doi: 10.3368/jhr.54.3.0517-8756R.