Teacher Choices: The Story Time Trial

Statistical Analysis Plan

Evaluator: National Foundation for Educational Research (NFER)

Principal investigator: Dr Ben Styles

Template last updated: August 2019

PROJECT TITLE	The Story Time Trial
DEVELOPER (INSTITUTION)	Education Endowment Foundation (EEF)
EVALUATOR (INSTITUTION)	National Foundation for Educational Research (NFER)
PRINCIPAL INVESTIGATOR(S)	Ben Styles
PROTOCOL AUTHORS	Maria Galvis, Palak Roy and Jennie Harland
SAP AUTHORS	Joana Andrade, Afrah Dirie and Palak Roy
TRIAL DESIGN	Two-armed randomised controlled trial (RCT) with randomisation at the school level
TRIAL TYPE	Feasibility of RCT
PUPIL AGE RANGE AND KEY STAGE	Ages 8-10 (Years 4 and 5)
NUMBER OF SCHOOLS	96
NUMBER OF PUPILS	8,448
PRIMARY OUTCOME MEASURE AND SOURCE	Pupils' comprehension skills (as measured by reading and listening comprehension of the text selected for the intervention, NFER's bespoke assessment)
SECONDARY OUTCOME MEASURE AND SOURCE	**Attainment** Pupils' wider reading comprehension skills (as measured by the reading comprehension of an unseen text, NFER's bespoke assessment) **Non-cognitive** Pupils' engagement in reading lessons, confidence in reading, liking of reading measured via a pupil survey using the scales from the PIRLS (Progress in International Reading Literacy Study) (Martin *et al.,* 2017)

SAP version history

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

Table of contents

Introduction

This Teacher Choices feasibility RCT aims to answer the research questions centred on the feasibility of the research method as well as the implementation, process and impact evaluation of the teacher choice.

This document presents the research questions from the protocol (Galvis et al., 2021), that are relevant to this SAP. Within the impact evaluation, the primary research question for this trial is:

1. What is the difference in pupils' text-aligned comprehension skills who received the GO! approach compared to those who received the STOP! approach? (RQ5a from the protocol). This measure will comprise the reading and listening comprehensions of the text selected for the intervention and therefore it is a text-aligned comprehension skill.

Within the impact evaluation, the secondary research questions for this trial are:

2. What is the difference in pupils' wider reading comprehension skills (of an unseen text) who received the GO! approach compared to those who received the STOP! approach? (RQ5a from the protocol)
3. What is the difference in pupils' engagement in reading lessons who received the GO! approach compared to those who received the STOP! approach? (RQ5b from the protocol)
4. What is the difference in pupils' confidence in reading who received the GO! approach compared to those who received the STOP! approach? (RQ5b from the protocol)
5. What is the difference in pupils' liking of reading who received the GO! approach compared to those who received the STOP! approach? (RQ5b from the protocol)

Additional analyses

There are a number of additional analyses that relate to the research questions of trial implementation rather than the impact evaluation. These are:

6. Is it possible to derive a suitable baseline measure by combining the scores of different commercial age-standardised English or reading assessments? (RQ3a from the protocol)

Another aim of this trial is to determine if a bespoke assessment that has not been trialled can be used as a valid and reliable outcome measure.

7. What psychometric properties does the bespoken assessment need to have in order to be used as a valid and reliable outcome measure for future trials? (RQ3b from the protocol)

Design overview

Trial design, including number of arms		Two-armed RCT
Unit of randomisation		School
Stratification variables		Not applicable
Primary outcome	variable	Text-aligned reading and comprehension score (TARC)
	measure (instrument, scale, source)	Reading and listening comprehensions of the text selected for the intervention, 0 to 35 marks, bespoke assessment developed by the NFER
Secondary outcome(s)	variable(s)	1. Wider reading comprehension score (WRC) 2. Engagement in reading lessons (ERL) 3. Student confidence in reading (SCR) 4. Student liking of reading (SLR)
	measure(s) (instrument, scale, source)	1. Reading comprehension of an unseen text, 0 to 16 marks, bespoke assessment developed by the NFER 2. PIRLS ERL scale, Pupil survey using existing scales from the PIRLS assessment 3. PIRLS SCR scale, Pupil survey using existing scales from the PIRLS assessment 4. PIRLS SLR scale, Pupil survey using existing scales from the PIRLS assessment

*There is no baseline for this trial

Randomisation

As per the protocol (Galvis et al., 2021), we have implemented a simple school-level randomisation design with no stratification. The R code used to perform the simple randomisation is included in Appendix 1 and all the calculations were performed in R 4.0.3. Overall, 96 schools were randomised equally across the two groups – 48 schools were randomised to each of the GO! And STOP! Approaches.

The participating schools were invited to nominate as many Year 4 and Year 5 classes as they saw fit to be included in the trial, and although the number of classes in a given year group or ratio of Year 4 to Year 5 classes in a school are not randomisation stratifiers; it is, nevertheless, desirable that these figures are balanced over the two randomisation groups. These distributions are displayed in Tables 1 below. Since randomisation, 10 schools have

withdrawn from the trial which leaves 86 schools retained in the trial- 42 in Go! and 44 in STOP!.

Table 1: Ratio of Year 4 to Year 5 classes across randomisation groups

Randomisation Group vs Y4 to Y5 ratio	1/2	2/3	3/4	1	2	Unknown or school only has mixed Y4/Y5 classes
GO!	1	1	1	38	1	6
STOP	4	0	0	38	1	5

Sample size calculations overview

		Protocol	Randomisation	
		OVERALL	OVERALL	FSM
Minimum Detectable Effect Size (MDES)*		0.204	0.208	0.239[1]
Intracluster correlations (ICCs)	level 2 (school)	0.12	0.12	0.12
Alpha		0.05	0.05	0.05
Power		0.8	0.8	0.8
One-sided or two-sided?		Two-sided	Two-sided	Two-sided
Average cluster size		90	88*	17*
Number of schools	STOP!	50	48	48
	GO!	50	48	48
	total	100	96	96
Number of pupils	STOP!	4,500	4,224	864
	GO!	4,500	4,224	864
	total	9,000	8,448	1,728

* Estimated

Given that the main aim of the trial is about the feasibility of running such a trial in primary schools, the target of 100 primary schools was decided to make the trial as big as feasible within the available timeframe and it was not governed by ability to detect differences between the randomisation groups. In the original trial design we had estimated that a sample of 100 schools with an average cluster size of 90 pupils (across years 4 and 5), an ICC of 0.12 (EEF, 2015), and a pre-post correlation of 0.73 (EEF, 2013) would equip the trial analysis with the power to detect a MDES of 0.15. However, in its present form the trial will include no baseline measures, which increases the MDES at protocol to 0.20 and 0.21 at randomisation.

[1] Similar to the overall MDES, the caveat about the deliberate underpowered analysis applies to the FSM subgroup too.

The power calculations were completed using a bespoke Excel spreadsheet and PowerUpR package in R.

Analysis

Before we discuss the impact of the outcome measures, we describe how the outcome measures are created for this trial.

Cognitive Outcomes (Primary and Secondary)

NFER's Centre for Assessment created a bespoke end-point assessment to measure the proximal outcomes of reading and listening comprehension of the selected book and the distal outcomes of reading comprehension of an unseen text. The structure of the end-assessment is briefly described in Table 2 below.

As seen in Table 2, the bespoken end-point assessment has a total of 51 one-mark items and is divided into four sections:

1. Global questions on the selected test (GQ) , with a total of four items
2. Listening comprehension on the selected test (LC), with a total of 20 items
3. Reading comprehension on the selected test (RC), with a total of 11 items
4. Wider reading comprehension based on an unseen text (WRC), with a total of 16 items

The primary outcome for the trial will be text-aligned reading and comprehension (TARC) as measured by the total score of the first three sections of the end-point assessment. The secondary outcome will be wider reading comprehension based on an unseen text (WRC) as measured by the total scores for the last section of the end-point assessment.

Table 2: Overview of the bespoke end-point assessment

Assessment section/outcome measures	Number of independent items	Total marks	Outcome measure
Global questions on the selected text (GQ)	4	4	Primary outcome (TARC)
Listening comprehension on the selected text (LC)	20	20	
Reading comprehension on the selected text (RC)	11	11	
Wider reading comprehension based on an unseen text (WRC)	16	16	Secondary outcome (WRC)

Non-Cognitive secondary outcomes

Three non-cognitive proximal outcomes will be derived from a survey filed by the pupils participating in the trial. The survey comprise items from the PIRLS 2016 student questionnaires and it is administered after the conclusion of the intervention but before the end-point assessment. The non-cognitive proximal outcomes correspond to the attitudes of students towards reading and will be measured by PIRLS scales constructed using item response theory (IRT), as described in Martin et al., 2017. The secondary outcome measure pupil engagement in reading lessons (ERL), their confidence in reading (SCR) and their enjoyment of reading (SLR). These outcomes are described in detail in the protocol and also in (Martin et al., 2017).

Psychometric evaluation of the bespoken assessment

As noted in the protocol, the end-point assessment did not undergo a formal assessment development process. Therefore, we will perform psychometric analysis of the assessment.

All the psychometric evaluation analysis described below will be run in R (version 4.1.2) or using customised SPSS and Mplus macros developed by psychometricians at NFER's Centre for Statistics.

Test statistics and score plots

We will report the following classical test theory summary statistics for the overall test as well as the TARC and WRC subscales:

1. Total number of respondents
2. Maximum and minimum scores
3. Mean score and standard deviation
4. Median score

We will also assess the reliability of the overall test and its two derived subscales by computing two reliability coefficients: Cronbach's alpha and Guttman's lambda 4 (Guttman, 1945). Although Cronbach's alpha is the industry standard for quantifying reliability it is also known to underestimate the true reliability of tests that are simultaneously measuring more than one construct (multidimensionality) (McNeish, 2018). Given that the bespoken assessment developed for the Story Time trial is measuring different content strands (e.g. reading and comprehension; and listening and comprehension) and also distal and proximal outcomes, we expect it to be multidimensional and as such, we will be considering a second reliability coefficient that does not tend to underestimate the reliability of multidimensional tests.

Although Guttman's lambda 4 is known to overestimate reliability for samples of less than 1000 observations and tests with a large number of items (Benton, 2015) it is, nevertheless, a reasonable choice in the context of the Story Time trial. Guttman's lambda 4 is given by:

$$\lambda 4 = max \frac{4.covariance(h1\ score, h2\ score)}{variance(total\ score)}$$

Where $h1$ and $h2$ are the halves of a split of the test or subscale items.

Both Cronbach's alpha (lambda 3) and Guttman's lambda 4 can be calculated by means of the function 'guttman' contained in the R package 'Lambda4' (Hunt, 2013).

In addition to the summary and reliability statistics we will also present, for both the test and derives subscales, score plots: histograms describing the distributions of test or subscale scores.

Item statistics

The following classical test theory statistics will be reported for each of the 51 one-mark items included in the bespoken assessment:

1. Facility or P-value: the percentage of pupils who answered the item correctly.
2. Percentage of omission: the percentage of pupils who skipped or did not answer the item.
3. Discrimination: the point-biserial correlation between the item and the total score of all the other items in the test.
4. Discrimination (subscale): the point-biserial correlation between the item and the total score of all the other items in the subscale the item is associated to (TARC or WRC).

Differential item functioning (DIF) analyses

Differential item functioning (DIF) analyses aim to determine if different groups within a population respond differently to particular items within a scale after controlling for group differences in the overall outcome being assessed. One of the possible approaches to this type of analysis, logistic regression DIF (Swaminathan and Rogers, 1990), models the likelihood of a pupil answering an item correctly via a logistic regression model with an interaction term:

$$\ln\left(\frac{p_j}{1-p_j}\right) = \beta_0 + \beta_1 \mathrm{score_j} + \beta_2 \mathrm{group_j} + \beta_3 \mathrm{score_j} * \mathrm{group_j} + \epsilon_j$$

Where p_j corresponds to the probability of pupil j answering the item correctly, $\mathrm{score_j}$ corresponds to their test or subscale score, and $\mathrm{group_j}$ to their group membership.

Based on the results of the logistic regression model, DIF is detected when pupils matched on test or subscale scores have a significantly different probability of answering the item correctly and thus differing logistic regression curves are observed. Conversely, if the curves for both groups are the same, then the item is unbiased and DIF is not present.

Logistic regression DIF analyses will be performed for the two outcomes in the bespoke assessment in order to enquire if DIF is present in terms of pupils' FSM eligibility, English as an additional language (EAL) status and gender. Separate analyses will be run for the subscale to which the item is associated (TARC or WRC), see Table 3.

Table 3: DIF analyses to be performed

	TARC subscale			WRC subscale		
	FSM	EAL	Gender	FSM	EAL	Gender
Items	Items 1 to 35			Items 36 to 51		

The logistic regression DIF analyses will be run on SPSS using macros developed by psychometricians at NFER's Centre for Statistics.

Confirmatory factor analyses

As discussed before, the bespoke endpoint assessment developed for the Story Time trial aims to assesses different content domains. As such, it is necessary to investigate if the underlying factor structure to the full test and its subscales, proximal and distal, correspond to the designs described in the outcomes section. We propose to do this by running the confirmatory factor analyses described in Table 4 below:

Table 4: Confirmatory factor analyses (CFA) to be performed

Analysis	Instrument	Factor	Associated items
CFA: two factors structure	Full test	TARC (proximal)	1 to 35
		WRC (distal)	36 to 51
CFA: four factors structure	Full test	GQ (global questions)	1 to 4
		LC (listening comprehension)	5 to 24
		RC (reading comprehension)	25 to 35
		WRC (wider reading comprehension)	36 to 51
		GQ (global questions)	1 to 4

CFA: three factors structure	TARC subscale (primary outcome)	LC (listening comprehension)	5 to 24
		RC (reading comprehension)	25 to 35
CFA: one factor structure	WRC subscale (secondary outcome I)	WRC (wider reading comprehension)	36 to 51

The results of the confirmatory factor analysis will be interpreted in conjunction with the values taken by the reliability coefficients derived for the full test and subscales in order to assess if the derived outcome measures are fit for use in the context of a randomised control trial. The confirmatory factor analysis will be run in R (version 4.1.2) using the package "lavaan" (Yves, 2012) or customised SPSS and Mplus macros developed at NFER's Centre for Statistics.

Primary and Secondary outcomes analyses

The primary and secondary analyses will follow the EEF 2018 guidelines, with intention-to-treat being assumed in all cases. As it was not possible to obtain a suitable baseline measure (such as Key stage 1) for assessing the pre-intervention reading attainment for all participating pupils, the evaluation team had re-designed the trial analyses to not include the baseline as a covariate.

The sample of participating pupils includes both Year 4 and Year 5 children being assessed with a common set of bespoken assessments and surveys. As it is to be expected that, as a group, Year 4 and Year 5 pupils perform differently in terms of the outcomes evaluated in the trial, the analyses models will all include a year group indicator to control for this fact. The year group indicator will be a dummy variable identifying Year 4 versus Year 5 pupils.

To account for the school-level randomisation design, all the models used in the analyses, both primary and secondary, will be two-level random intercepts models (pupil and school).

Primary outcome analysis

The primary analysis will compare the impact on pupils' proximal listening and comprehension skills of teachers adopting one of the two reading approaches, STOP! or GO!. For this purpose we will fit a model with text-aligned reading and comprehension, as measured by TARC score, as the dependent variable. As discussed above, the primary outcome model will not include a baseline covariate, but we will control for year group by means of a dummy variable (Year 4 versus Year 5).

The pupil-level regression model is given by:

$$\mathrm{TARC}_{ij} = \beta_{0j} + \beta_1 \mathrm{STOP!}_j + \beta_2 \mathrm{YG}_{ij} + \epsilon_{ij}$$

Where TARC_{ij} is the text-aligned reading and comprehension score of pupil i in school j, β_{0j} is the intercept in school j, $\mathrm{STOP!}_j$ is a pupil-level dummy indicator on whether the pupil was randomised approach was STOP!, and YG_{ij} is the year group of pupil i in school j.

The model will be run in R (version 4.1.2) using the package 'nlme' (Pinheiro, et al., 2021).

Secondary outcome analysis

Secondary analysis will compare the impact of the adopting one of the trial's reading approaches on pupils' distal listening and comprehension skills (as measured by WRC score). In accordance with the EEF 2018 guidelines, the specification of analysis model will be in the same mould as the primary analysis:

$$\mathrm{WRC_{ij}} = \beta_{0j} + \beta_1 \mathrm{STOP!_j} + \beta_2 \mathrm{YG}_{ij} + \epsilon_{ij}$$

Where $\mathrm{WRC_{ij}}$ is the wider reading and comprehension score of pupil i in school j, β_{0j} is the intercept in school j, $\mathrm{STOP!_j}$ is the pupil-level STOP! indicator, and YG_{ij} is the year group of pupil i in school j. The error associated with the i^{th} pupil in school j is given by ϵ_{ij}.

Similarly, the comparison of the impact of the reading choice on three non-cognitive pupil outcomes will be described by the following two-level regression models:

$$\mathrm{ERL_{ij}} = \beta_{0j} + \beta_1 \mathrm{STOP!_j} + \beta_2 \mathrm{YG}_{ij} + \epsilon_{ij}$$

$$\mathrm{SCR_{ij}} = \beta_{0j} + \beta_1 \mathrm{STOP!_j} + \beta_2 \mathrm{YG}_{ij} + \epsilon_{ij}$$

$$\mathrm{SLR_{ij}} = \beta_{0j} + \beta_1 \mathrm{STOP!_j} + \beta_2 \mathrm{YG}_{ij} + \epsilon_{ij}$$

Where, $\mathrm{ERL_{ij}}$ is the engagement in reading lessons score of pupil i in school j, $\mathrm{SCR_{ij}}$ is the confidence in reading score of pupil i in school j, and $\mathrm{SLR_{ij}}$ is the enjoyment of reading score of pupil i in school j. The covariates of the non-cognitive secondary outcomes models are as specified for the primary analysis.

The secondary models will be run in R (version 4.1.2) using the package 'nlme'.

Subgroup analyses

Subgroup analyses will be performed to assess the differential impact of the choice of reading approach on the primary outcomes of two groups of pupils: FSM eligible pupils and pupils with English as an additional language (EAL)[2].

We will approach the subgroup analyses in two distinct ways: we will run models with interaction terms (i.e. models that include both the subgroup indicator and the product of the subgroup indicator and randomised reading approach), and we will run separate primary outcome models on just the subgroups of interests. Both approaches conform to the EEF 2018 guidelines and underpowered subgroup analyses will be reported as exploratory.

The multilevel level random intercepts model with interaction terms for the FSM subgroup analysis will be given by:

$$\mathrm{TARC_{ij}} = \beta_{0j} + \beta_1 \mathrm{STOP!_j} + \beta_2 \mathrm{YG}_{ij} + \beta_3 \mathrm{FSM}_{ij} + \beta_4 \mathrm{STOP!_j} * \mathrm{FSM}_{ij} + \epsilon_{ij}$$

With FSM_{ij} being a dummy variable for pupil i in school j's FSM eligibility status and the remaining variables as described in the Primary Outcome Analysis subsection above.

And the model for the EAL subgroup analysis will be specified as:

$$\mathrm{TARC_{ij}} = \beta_{0j} + \beta_1 \mathrm{STOP!_j} + \beta_2 \mathrm{YG}_{ij} + \beta_3 \mathrm{EAL}_{ij} + \beta_4 \mathrm{STOP!_j} * \mathrm{EAL}_{ij} + \epsilon_{ij}$$

Where EAL_{ij} is a dummy variable for pupil i in school j's EAL status and the remaining variables as described in the Primary Outcome Analysis subsection above.

[2] Variables FSM and EAL will be collected directly from schools.

Additional analyses

Derivation of a baseline measure

As discussed in the protocol, it was not possible to collect KS1 AREs in 2020 and, as such, this trial will not include a baseline measure of reading and comprehension skills. Several alternative baseline measures to KS1 ARES were considered and discussed, and these discussions motivated a research question (RQ3a from protocol): Is it possible to derive a suitable baseline measure by combining the scores of different commercial age-standardised reading and English assessments?

Some of the schools participating in the trial ran English or reading assessments in the spring or summer term of 2021 and agreed to share their results. We will collect the assessment scores directly from four commercial providers[3] that created tests used by trial schools in 2021 and use them to derive a baseline measure for both primary and secondary outcomes.

As working hypotheses, we will assume the following:

1. Item-level data will not be collected, just total scores, but we will have access to information on the characteristics of the different assessments, including on the representativeness of the sample on which each assessment was trialled.
2. The providers will disclose the means and standard deviations of their assessment scores.
3. The commercial tests used in the spring or summer terms of 2021 to evaluate the reading and comprehension skills of children in the same year group, Year 3 and Year 4, were of similar difficulty and measured the same or very similar constructs.

Assuming the providers produced assessments that were developed in samples that are representative of the same population, we can link the tests for the same year via their z-scores (calculated using the means and standard deviations reported by the providers, not the individual trial samples of pupils who sat the same test). Under the reasonable assumption that reading and comprehension ability follows an approximately normal distribution in the two year groups computing z-scores matches the scores of the different tests to a year-group standardised common scale of measurement of reading and comprehension skills[4].

Being a year-group standardised measure, the baseline we defined needs to be interpreted considering the pupils' cohort (year 4 or year 5) while the endpoint bespoke assessment created by the evaluation team has been designed to apply similarly to both cohorts without adjusting for year group. Given the young age of the pupils taking part in the trial, we would expect older pupils in year 5 to generally do better than their younger counterparts in year 4, and a poor correlation between the year-group standardised baseline measure and the non-standardised primary outcome measure.

A possible way to increase the pre-post correlation and the primary analysis model's explained variance is to standardise the dependent variable, primary outcome TARC scores,

[3] The four test providers being Renaissance Learning, Rising Stars, NFER and GL Assessment.
[4] Pupil scores on the common reading and comprehension measurement scale correspond to the deviation of a pupil's reading and comprehension ability from the mean ability of their year group measured in standard deviations.

by year group. For this analysis set up, which we will denote by SU1, the primary model is given by:

$$\text{standTARC}_{ij} = \beta_{0j} + \beta_1 \text{STOP!}_j + \beta_2 \text{baseline}_{ij} + \beta_3 \text{YG}_{ij} + \epsilon_{ij}$$

Where standTARC_{ij} and baseline_{ij} are the standardised TARC and baseline scores defined above for pupil i in school j, and the remaining covariates of the model are as defined in the primary analysis section.

On the other hand, it should also be noted that even for children as young as the ones in the trial sample there is a degree of overlap in reading and comprehension skills of key stage 2 pupils in consecutive years. We expect that pupils at the top range of ability for their year group cohort will do well in a common assessment, and, conversely, that pupils at the lower range of ability for their year group cohort will perform poorly. Following this rationale, it can be argued that a degree of correlation between the primary outcome TARC scores and a year group standardised baseline measure is to be expected. To evaluate if a trial design that includes a non-standardised outcome and standardised baseline measure will work well in the context of a randomised control trial we will consider the following analysis set up, SU2, and the primary model:

$$\text{TARC}_{ij} = \beta_{0j} + \beta_1 \text{STOP!}_j + \beta_2 \text{baseline}_{ij} + \beta_3 \text{YG}_{ij} + \epsilon_{ij}$$

This model is identical to the trial's primary analysis model except for the inclusion of the extra covariate baseline_{ij} which corresponds to the baseline year group standardised baseline scores for pupil i in school j,

We will implement SU1 and SU2 in the subsample of pupils for which schools reading assessment data collected in the summer of 2021 is available and compare their results in terms of:

1. Pre-post correlation between outcome and baseline measures
2. ICCs
3. MDES

Imbalance at baseline

To assess imbalance between the STOP! And GO! at baseline we will produce cross-tabulations of background characteristics of the schools in the sample. We will examine the following school-level background characteristics:

1. School attainment
2. Proportion of FSM pupils
3. Urban or rural
4. Type of school governance
5. Latest Ofsted rating

To run this analysis, we will link the schools taking part in the trial to the relevant information contained on the most up to date edition of NFER's registry of schools.

We will also assess imbalance at baseline in terms in terms of number of pupils in Year 4 and Year 5, as well as the number and proportion of Year 4, 5 and mixed classes in the sample of schools.

Missing data

The analysis model only includes two covariates, randomised reading approach and year group, whose values are known for all the pupils participating in the trial and the only possible source of missingness in the trial's data is the dependent variable. As such, we will not investigate missingness in terms of any variables other than the primary outcome.

Moreover, the main aim of the trial is about the feasibility of running such a trial in primary schools and the sample size was decided to make the trial as big as feasible rather than the statistical power of the analysis. Hence, we will report levels of missingness in the primary outcome variable in a CONSORT flow diagram along with reasons of missing data (where possible), but we will not investigate patterns of missingness.

Compliance

Compliance will be measured using information collected via the online teacher feedback survey. The dosage measure for compliance will be total amount of time (in minutes) that the teacher used an allocated reading aloud approach. This will include time spent in reading aloud whilst completing the selected book and time spent while reading aloud any other book using the allocated approach for any remaining sessions within the intervention period. Compliance will c and will not form part of the impact analysis.

Intra-cluster correlations (ICCs)

ICCs will be estimated from the variance of the random intercept and residual variance of the analysis multi-level models by means of the formula:

$$ICC = \frac{\sigma^2_{school}}{\sigma^2_{school} + \sigma^2_{residuals}}$$

Pre-test ICCs will also be computed in the context of the additional analyses described above. In this cases we will be considering random intercepts two-level models, school and pupil, with no covariates.

Effect size calculation

As per the EEF 2018 guidelines, the results of the analyses of continuous outcomes by means of multi-level regression models will be reported as Hedges' g. The effect size will be calculated according to the formula

$$\boldsymbol{g} = \frac{\bar{\boldsymbol{o}}\boldsymbol{s}! \; - \; \bar{\boldsymbol{o}}\boldsymbol{g}!}{\boldsymbol{s}^*}$$

The numerator $\bar{\boldsymbol{o}}\boldsymbol{s}! - \bar{\boldsymbol{o}}\boldsymbol{g}!$ is the difference between the STOP! and GO! group in terms of the mean value of the outcome being assessed, and the denominator $\boldsymbol{s}^*$ is the pooled standard deviation of the outcome and is given by the formula

$$s^* = \sqrt{\frac{(N_{s!} - 1)s^2_{s!+}(N_{g!} - 1)s^2_{g!}}{N_{s!} + N_{g!} - 2}}$$

with $N_{s!}$ and $N_{g!}$ being the number of elements in the STOP! and GO! groups, and $s_{s!}$ and $s_{g!}$ the standard deviations of the outcome measured in the intervention and control groups

The numerator for the effect size calculations can be calculated as the coefficients of the randomisation group from the regression models, and the denominator as the unconditional variance from the corresponding models without covariates. The effect sizes thus computed are equivalent to Hedges'g.

Confidence intervals for each effect size will be computed by multiplying the standard errors of the intervention group by the left-tailed inverse of the Student's t-distribution with a probability of 2.5% and the number of degrees of freedom associated to the size of the sample. The confidence intervals for the standard errors will be converted to effect size confidence intervals using the same formula as the effect sizes themselves.

References

Benton, T., 2015. *An empirical assessment of Guttman's Lambda 4 reliability coefficient*. [pdf] Available at: <141299-an-empirical-assessment-of-guttman-s-lambda-4-reliability-coefficient.pdf (cambridgeassessment.org.uk)> [Accessed 10 March 2022].

Department for Education (DfE), 2021. *Free school meals: Autumn term.* [online] Available at: <https://explore-education-statistics.service.gov.uk/find-statistics/free-school-meals-autumn-term/2020-21-autumn-term> [Accessed 10 March 2022].

Education Endowment Fund (EEF), 2013. *Pre-testing in EEF evaluations.* [pdf] Available at: <https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Protocol_or_SAP/Pre-testing_paper.pdf> [Accessed 10 March 2022].

Education Endowment Foundation (EEF), 2015. *Intra-cluster correlation coefficients*. [pdf] Available at: <ICC_2015.pdf (educationendowmentfoundation.org.uk)> [Accessed 10 March 2022].

Galvis, M., Roy, P., Harland, J., and Lord, P., 2021. *Teacher choices: The story time trial evaluation protocol*. [pdf] Available at: <https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/Trial_2_Evaluation_Protocol.pdf> [Accessed 10 March 2022].

Guttman, L., 1945. A basis for analysing test-retest reliability. *Psychometrika*, [e-journal] 10, pp.255-282. Available through: Springer Link website <A basis for analyzing test-retest reliability | SpringerLink> [Accessed 10 March 2022].

Martin, M. O., Mullis, I. V. S., and Hooper, M., Liqun, Y., Foy, P., Fishbein, B. and Liu, J., 2016. *Creating and Interpreting the PIRLS 2016 Context Questionnaire Scales.* [pdf] Available at: <https://timssandpirls.bc.edu/publications/pirls/2016-methods/P16_MP_Chap14_Context_Questionnaire_Scales.pdf [Accessed 10 March 2022].

McNeish, D., 2018. Thanks coefficient alpha, we'll take it from here. *Psychological Methods,* [e-journal] *23*(3). Available at: <Thanks Coefficient Alpha, We'll Take it From Here (researchgate.net)> [Accessed 15 December 2021].

Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team, 2021. *nlme: Linear and Nonlinear Mixed Effects Models. R (3.1-155)*. [computer program] Available at: < https://CRAN.R-project.org/package=nlme> [Accessed 10 March 2022].

Swaminathan, H. and Rogers, J., 1990. Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, [e-journal] 27(4), pp.361-370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x.

Tyler Hunt, 2013. *Lambda4*: *Collection of Internal Consistency Reliability Coefficients. R* (3.0.). [computer program] Available at: <https://cran.r-project.org/package=Lambda4> [Accessed 10 March 2022].

Yves R., 2012. lavaan: An R Package for Structural Equation Modelling. *Journal of Statistical Software*, [e-journal] 48(2), pp.1-36. 10.18637/jss.v048.i02

Appendix 1: School level randomisation

```
###5. Load data

filename=dir()[grep("EETC2 Randomisation Info",dir())]

RandFile=read.csv(filename,stringsAsFactors = FALSE)


###I set the seed to the day I'm writing the code

set.seed(….)


### Randomise to NFER No (unique school identifier)


####Sanity check: there are no repeated NFER numbers and all the cases have one assigned

###Number of NFERNos matches number of school names: All good!

sum(is.na(RandFile$nfer))==0

sum(duplicated(RandFile$nfer))==0

length(unique(RandFile$school_name))==length(unique(RandFile$NFE
R_no))

###Keep original row and column order

###In the end of the randomisation I'll want an extra randomisation group: which I'll call Rand_S (_S identifies the randomisation) is at school level

Colorder=c(colnames(RandFile),"Rand_S")

RandFile$original=1:(nrow (RandFile))

###Order the rows of RandFile by NFER no

RandFile=RandFile[order(RandFile$NFER_no),]


###And randomly reorder them
```

```
Aux=sample(1:nrow(RandFile))
RandFile=RandFile[aux,]
### Assign randomisation group:GO! and STOP!
RandFile$aux=1:(nrow(RandFile))
RandFile$aux=(RandFile$aux%%2)+1

Randgroup=data.frame(Rand_S=c("GO!","STOP!"))
Randgroup$aux=sample(1:2)
RandFile=merge(RandFile,Randgroup)
##Returning the data frame to its original row order
RandFile=RandFile[order(RandFile$original),]
###And putting the columns back in order
RandFile=RandFile[,colorder]
```