REACH Primary
Statistical Analysis Plan
Evaluator (institution): Sheffield Hallam University
Principal investigator(s): Prof Mike Coldwell

Education Endowment Foundation

Template last updated: August 2019

PROJECT TITLE[1]	READING WITH COMPREHENSION (REACH) Primary, a two-arm cluster randomised trial
DEVELOPER (INSTITUTION)	University of Leeds
EVALUATOR (INSTITUTION)	Sheffield Hallam University
PRINCIPAL INVESTIGATOR(S)	Professor Mike Coldwell
SAP AUTHOR(S)	Dr Martin Culliney, Sean Demack, Dr Josephine Booth
TRIAL DESIGN	Two-arm multisite cluster randomised controlled trial with random allocation at school level
TRIAL TYPE	Efficacy
PUPIL AGE RANGE AND KEY STAGE	Year 3 (Age 7-8, KS2)
NUMBER OF SCHOOLS	80
NUMBER OF PUPILS	800
PRIMARY OUTCOME MEASURE AND SOURCE	Pearson Wechsler Individual Achievement Test 3rd edition, Reading Comprehension subtest (WIAT III)
SECONDARY OUTCOME MEASURE AND SOURCE	GL Diagnostic Test of Word Reading Processes (DTWRP) Pearson Understanding Spoken Paragraphs subscale of the Clinical Evaluation of Language Fundamentals (CELF -5)
TRIAL REGISTRATION	http://www.isrctn.com/ISRCTN15145068

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0	21/05/2020	*N/A*

Table of contents

Introduction

REAding with CompreHension Primary (REACH Primary) is a targeted intervention for struggling readers that comprises two strands: Reading Intervention (RI) sessions, delivered twice per week, and Oral Language Intervention (OL) sessions, delivered weekly.

This statistical analysis plan outlines the impact analyses, discusses the RCT design and provides sample size calculations. It also addresses the primary and secondary outcome analyses, sub-group analysis, the handling of missing data and noncompliance issues, and finally effect size calculations.

Design overview

Table 1 presents an overview of the trial design.

Table 1: Design overview

Trial design, including number of arms		Two-arm, cluster randomised [3 level clustered trial blocked by geographical area; a 4-level multisite CRT]
Unit of randomisation		School
Stratification variables (if applicable)		Geographical area, school level mean KS1 reading score for participating pupils
Primary outcome	variable	Wechsler Individual Achievement Test 3rd edition, Reading Comprehension subtest
	measure (instrument, scale, source)	Raw scores (scale 0-42)
Secondary outcome(s)	variable(s)	Diagnostic Test of Word Reading Processes (decoding of words, non-words and exception words) CELF-5 'Understanding Spoken Paragraphs' (scale 0-20)
	measure(s) (instrument, scale, source)	DTWRP: standard age scores (average score is 100) CELF-5: paper version, raw scores
Baseline for primary outcome	**variable**	KS1 Reading
	measure (instrument, scale, source)	KS1 Reading, raw scores combined for both papers (scored 0-40), data collected directly from schools prior to randomisation
Baseline for secondary outcome	**variable**	KS1 Reading
	measure (instrument, scale, source)	KS1 Reading, as above

Sample size calculations overview

Table 2 presents MDES and sample sizes for the protocol and randomisation stages. Protocol estimates are based on a predicted number of pupils per TA (10) and TAs per school (2). For the randomisation stage, data was collected from schools. This is a 3-level clustered RCT stratified by geographical area, labelled by Spybrook et al. (2016) as a 4-level multisite clustered RCT (MSCRT). Please see Appendix I for the MDES formula used. At both protocol and randomisation, the estimated MDES for the primary outcome is 0.24. Both are lower than the +0.33 sd effect size found in the previous evaluation of REACH (in secondary schools). For the Free School Meals (FSM) subgroup analyses the MDES estimate is 0.29 at both protocol and randomisation.

Table 2: Sample size calculations

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
Minimum Detectable Effect Size (MDES)		0.24	0.29	0.24	0.29
Pre-test/ post-test correlations	level 1 (pupil)	0.74	0.74	0.74	0.74
	level 2 (TA)	0.60	0.60	0.60	0.60
	level 3 (school)	0.60	0.60	0.60	0.60
Intracluster correlations (ICCs)	level 2 (TA)	0.02	0.02	0.02	0.02
	level 3 (school)	0.14	0.14	0.14	0.14
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.80	0.80	0.8
One-sided or two-sided?		2	2	2	2
Average cluster size		5 pupils per TA, 10 per school	2 FSM pupils per TA, 4 per school	5 pupils per TA, 10 per school	2 FSM pupils per TA, 4 per school
Number of schools	intervention	40	40	40	40
	control	40	40	39	39
	total	80	80	79	79
Number of TAs	intervention	80	80	81	81
	control	80	80	72	72
	total	160	160	153	153
Number of pupils	intervention	400	160	391	138
	control	400	160	389	251
	total	800	320	780	255

The MDES estimates shown in Table 2 assume a 3-level CRT design stratified by geographical area (defined as a 4-level MSCRT in Spybrook et al., 2016). The inclusion of a TA level in the design brings advantages in terms of the precision in measuring compliance

to REACH training (see below). Including the TA level allows compliance to be related to a specific TA and any within-school between-TA variance in compliance to REACH can be captured. However, the viability of a TA level is unknown at this point in time. If the TA level is not viable, due to substantial movement of pupils between TAs or poor completion of TA logs linking pupils to TAs during delivery, the multilevel design will become a 2-level CRT blocked by geographical area (a 3-level MSCRT).

If the TA level is dropped, a 3-level MSCRT design assuming 10 pupils per school (4 FSM pupils per school) results in an MDES estimate of 0.24 standard deviations overall and 0.28 for the FSM subsample. The similarity in MDES estimates for the 4 level and 3 level MSCRT designs relates to our assumption that clustering at the TA level is relatively weak (TA level ICC assumed to be 0.02).

Analysis

The impact analysis will be structured to address the following three research questions:

RQ1. What is the impact of REACH Primary on pupil reading comprehension ability, as measured by the WIAT III Reading Comprehension subtest?
RQ2. What is the impact of REACH Primary on pupil word recognition and decoding ability, as measured by GL DTWRP?
RQ3. What is the impact of REACH Primary on pupil language comprehension, as measured by the 'Understanding Spoken Paragraphs' module of the Pearson CELF-5 test?

A multilevel approach will be taken, with pupils clustered into TAs and TAs clustered into schools (3-level random intercepts multilevel models). Multilevel linear regression models will be constructed for the primary outcome, which is the Wechsler Individual Achievement Test 3rd edition, Reading Comprehension subtest (henceforth WIAT III), using Stata. KS1 Reading will be used as the baseline covariate. The first model will only include the school level group identifier (an outcome only model). The second model will also include KS1 Reading as a covariate at the pupil, TA and school level, using data collected directly from schools. The final model will also include the two variables used within the stratified randomisation (geographical hub area, school level mean KS1 Reading). This final model will be used for the headline ITT impact analysis for the WIAT III primary outcome, addressing Research Question 1. All models are summarised in Table 3. Further detail on variance decomposition and centring of covariates at school, TA and pupil levels can be found in Appendix II.

For each model, the coefficient of the school level dummy variable is used to distinguish 'intervention group' pupils, at schools who will receive the REACH Primary programme, from 'control group' pupils. This coefficient will be converted into Hedges' g effect size statistics with 95% confidence intervals. Appendix II provides more technical detail on the multilevel model that will be used for the ITT (headline) analyses of the WIAT III primary outcome and how the Hedges' g effect size statistic will be calculated.

The two secondary outcomes are the Diagnostic Test of Word Reading Processes (DTWRP) and the Understanding Spoken Paragraphs subscale of the Clinical Evaluation of Language Fundamentals (CELF-5). These measures will be subject to the same analysis process as the primary outcome, although analyses will be exploratory and will not be used determine

the efficacy of the programme. Analysis of these two secondary outcomes will address Research Questions 2 and 3 respectively.

Table 3: Analysis models

Analysis and Sample	*Level 1 (pupil) Variables*	*Level 2 (TA) Variables*	*Level 3 (school) Variables*
ITT sample	-	-	• Group (1=intervention; 0=control)
ITT sample	KS1 Reading (TA centred)	Mean KS1 Reading (school centred)	• Group (1=intervention; 0=control); • Mean KS1 Reading (Grand mean centred)
Final (headline) Analysis ITT sample	KS2 Reading (TA centred)	Mean KS1 Reading (school centred)	• Group (1=intervention; 0=control); • Mean KS1 Reading (Grand mean centred) Stratification variables: • Geographical hub area

Primary outcome: WIAT III Reading Comprehension subtest. Secondary outcome 1: Diagnostic Test of Word Reading Processes (DTWRP).Secondary outcome 2: Understanding Spoken Paragraphs subscale of the Clinical Evaluation of Language Fundamentals (CELF -5). All scheduled for summer 2020.

Subgroup analyses

Separate subgroup analyses on Free School Meals (FSM), English as an Additional Language (EAL) and Special Educational Needs (SEN) pupils will be conducted. Relevant indicators for participating pupils were supplied by schools prior to randomisation, although these will also be collected from NPD for use in the final analysis. There is very little missing data for the subgroup indicators as supplied by schools. There are no missing values for FSM, two for SEN and 11 for EAL. The FSM variable used in the final analysis will be the NPD EVERFSM6 indicator.

In line with 2018 EEF analysis guidance, these analyses will be undertaken in two stages. First, three models will be constructed that include two additional variables; the pupil level FSM/EAL/SEN binary identifier and an interaction between the identifier and group membership (FSM*group). These models will examine whether there is evidence that the REACH intervention had a differential impact on attainment for pupil subgroups (defined by FSM, EAL & SEN). A statistically significant interaction would provide evidence of differential impact. For the EAL and SEN analyses, if the interaction term is found to be statistically significant (two tailed, $p<0.05$), follow-on subgroup analyses will be undertaken. In line with EEF analysis guidance, follow-on analyses of FSM and non-FSM pupil subsamples will be undertaken regardless of the findings from the analyses of interaction effects. All subgroup analyses are purely exploratory.

Additional analyses

The ITT analysis will be repeated but with the KS1 Reading covariate only included at the pupil level. The estimated effect size from this model will be compared with the ITT estimate. Additionally, the explanatory power of the KS1 Reading covariate at school, TA and pupil

levels will be examined through comparing the explanatory power provided by a model with only a pupil level variable with a model including the KS1 Reading covariate at all levels.

Longitudinal follow-up analyses

Relevant outcomes for longitudinal analysis would be KS2 Reading and KS4 English. These analyses would be conducted according to the principles outlined in EEF Longitudinal Analyses Guidance.

Imbalance at baseline

Table 4 displays descriptive statistics comparing the intervention and control groups at baseline. The geographical distribution of participating schools is well balanced as each hub area was randomised separately to ensure this. This approach was taken so that the number of schools attending training in each area was manageable for the delivery team.

Intervention and control schools show very similar numbers of disadvantaged pupils at both KS1 and KS2. The percentage of pupils reaching the expected standard in reading is greater in control schools (68.4% compared to 65.4%), as is the percentage of pupils reaching the expected standard in reading, writing and maths (57.4% compared to 55.2%). These figures are not derived from the pupils sampled in this study, but are taken from DfE schools comparison data.

There is also imbalance at baseline in the pre-test KS1 Reading measure which was supplied by schools for each pupil as a condition for participating in the study. The control group again has higher scores on this test (11.3 compared to 10.6), which is scored on a scale from 0-40. This gives an effect size, calculated by dividing the mean difference by the pooled standard deviation, of -0.09.

Table 4: Imbalance at baseline

	Baseline ($N_{Schools}$=79)		Analysis ($N_{Schools}$=)	
	Intervention group (N=40)	Control group (N=39)	Intervention group (N=)	Control group (N=)
School level (categorical)	% (n)	% (n)	% (n)	% (n)
Hub Area				
Lincs	17.5%(7)	15.4%(6)		
North East	17.5% (7)	17.9%(7)		
North West	20%(8)	23.1%(9)		
South Yorkshire	25%(10)	25.6%(10)		
West Yorkshire	20% (8)	17.9%(7)		
OFSTED Grades				
Outstanding	10.1%(5)	7.7%(3)		
Good	72.2%(27)	76.9%(30)		
Requires Improvement	15.2%(7)	12.8%(5)		
Inadequate	0%(0)	2.6%(1)		
Missing	1.3%(1)	0%(0)		
School level (continuous)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
KS1_Reading score	10.8(4.7)	11.1(4.55)		
Total number of pupils (including part-time pupils)	467(103)	482(142)		
Percentage of key stage 2 disadvantaged pupils	43.7(18.81)	43.8(20.43)		
Cohort level key stage 1 average points score	15.2(1.2)	15.3(1.27)		
Percentage of eligible pupils with EAL	19.9(29.29)	20.3(29.69)		
Percentage of eligible pupils with SEN	2.2(3.09)	1.6(1.69)		
Percentage of pupils reaching the expected standard in reading	65.4(16.36)	68.4(13.24)		
Percentage of pupils reaching the expected standard in reading, writing and maths	55.2(17.26)	57.4(14.15)		
Pupil level (continuous)	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)
Pre-test score	10.6 (7.32)	11.3 (7.24)		
effect size (Int-Cont)	-0.09			

Missing data

There were no missing data at baseline (KS1 reading scores, group membership, geographical hub), therefore the only possible missing data will be found in the primary outcome. The reasons for any missing data (such as school/pupil withdrawal) will be recorded and summarised in the final report.

In the instance of any missing outcome data, the (complete) baseline and ITT samples will be compared across all ITT variables and additional variables shown in Table 4 above.

In the instance of over 5% of missing outcome data, as part of the follow-on analyses a multilevel logistic regression model with a binary outcome identifying when outcome data is missing (=1) or not (=0) will be constructed. The ITT variables and additional school level variables shown in Table 4 will be used to identify whether the missing outcome data can be assumed to be missing at random.

If none of the explanatory variables are found to account for a statistically significant amount of variation in the missing data outcome, we will cautiously assume that the data is missing at random[2]. This leads to cautiously concluding that the ITT estimate is not biased due to missing data.

If one or more explanatory variables are found to account for a statistically significant amount of variation in the missing data outcome we would undertake a sensitivity analysis to repeat the ITT analysis with these variables included. The potential bias introduced by missing outcome data on the ITT estimate will be illustrated by comparing the estimated ITT effect size with the effect size estimated from the ITT model including the additional variables.

Compliance and dosage

Compliance will be measured at the TA level, through TA training. Dosage will be measured at the pupil level, through intervention delivery. Full details are provided in Table 5. The TA compliance and pupil dosage measures will be combined to create overall minimal and optimal compliance indicators at the pupil level.

Compliance with the TA training will be assessed according to three criteria that were developed in collaboration with the developers at the University of Leeds:

Criterion 1: The intervention entails three face-to-face training days, and attendance will be used as a compliance measure. If a TA does miss a session, compliance can only be obtained through a visit by a member of the delivery team.[3] Where this does not happen, participating TAs are encouraged to seek input from colleagues who have attended the training but this is not deemed an adequate substitute for personally attending and will not count as attendance for the purpose of calculating compliance. Completion of REACH Primary training is the sole criterion for defining minimal compliance for TAs.

Criterion 2: Attendance of four online seminars (5 in total). Five online seminars for participating TAs are offered during the intervention; at least four of these must be attended, as indicated by seminar attendance lists and video watching statistics, for optimal compliance.

Criterion 3: A minimum of five of the eight gap tasks must be completed for optimal compliance. TAs must complete gap tasks within 14 days of the expected completion date.

[2] We must be cautious because we are limited to the variables included in the missing data logistic regression model. There will always be the potential of unmeasured variables.
[3] All schools who missed a training session were offered a visit.

To verify completion, a combination of statistics on who has viewed the online material and those who have answered a short 'quiz' at the end of each session is used.

The three TA level criteria will be drawn together to construct two TA level binary variables; the first defining minimal compliance (1=completing REACH training; 0=not completing) and the second defining optimal compliance (1=completing REACH training, online seminars and gap tasks).

Dosage, assessed at pupil level, will be measured by the number of intervention sessions delivered. REACH Primary comprises two distinct components: Reading (38 sessions) and the Comprehension (19 sessions). The total pupil level dosage will be calculated using equation 1.1:

Equation 1.1: $$Dosage_{pupil} = \frac{\left[\frac{Reading\ Sessions}{38} + \frac{Comprehension\ Sessions}{19}\right]}{2}$$

Pupils must complete 34 Reading Intervention sessions and 17 Comprehension sessions (in addition to the sessions scheduled for the first week of the intervention) to be considered compliant. This will enable the construction of a single binary pupil level measure of compliance (1=completing 89%+ Reading AND 89%+ Comprehension REACH sessions). Minimal compliance is completing the equivalent of the first term's material, 47% of the RI and C sessions (nine Comprehension sessions and 18 Reading Intervention sessions).

The recommended time for each session is 30 minutes. Each consists of a structured activity schedule. It is therefore expected that TAs use the full amount of time allocated for the sessions so that it is possible to cover all of the necessary content. The evaluators are collecting data from TAs on the duration of each session. Those lasting for less than 25 minutes will not be counted as complete, as this is considered by the developers to be the minimum length of time required to cover all constituent elements of the sessions.

Table 5: Compliance indicators, TA and pupil level.

Activity	Description	Data source	Measurement (minimal)	Measurement (optimal)
TA Training (TA level)				
TA Face to Face Training	Content covered through attending sessions or a visit from a researcher	Attendance list	All three sessions attended	All three sessions attended
TA Gap Tasks	Completed within 14 days of expected date	Short test at the end proves completion	N/A	Five of the eight tasks must be completed
Online Seminars	Sessions attended or video summaries watched after the event	Attendance lists in online seminar; video watching statistics	N/A	Four of the five sessions must be completed
Intervention delivery (pupil level)				
Reading Intervention	Pupil completes weekly session, which lasts at least 25 minutes	TA log	18 of 38 sessions completed, plus week 1	34 of 38 sessions completed, plus week 1
Comprehension Sessions	Four core activities completed per session	TA log	9 of 19 sessions must be completed	17 of 19 sessions must be completed

The TA and pupil level dimensions of compliance will be drawn together to create the overall minimum and optimal compliance measures as follows:

TA minimal compliance (0 or 1) * pupil minimal dose (0 or 1) = Minimal compliance (0 or 1)

TA optimal compliance (0 or 1) * pupil optimal dose (0 or 1) = Optimal compliance (0 or 1)

These variables will be used to estimate the Complier Average Causal Effect (CACE). The purpose of the Complier Average Causal Effect (CACE) analysis is to estimate the impact of REACH for pupils deemed to have 'complied' with the intervention.

CACE will be estimated using two stage least squares (2SLS) regression (Gerber & Green, 2012). The first stage will model the pupil-level compliance variables (first minimal then optimal) using the same explanatory variables listed in Table 3 for the headline ITT analyses along with additional school level items that are available via the school census as included in Table 4. This will be a multilevel logistic regression model used to generate predicted minimal/optimal compliance (1 or 0) for use in the second stage model. The second stage models will use predicted compliance in place of the group identifier variable in the ITT analyses specified above to generate the CACE estimates for REACH Primary.

Two CACE estimates will be calculated.

- First, using the predicted minimal compliance variable; 1=pupils who attended 47%+ Reading and 47%+ Comprehension sessions led by a TA who completed the REACH training; 0=control pupils plus pupils in intervention schools who attended

<47% Reading and <47% Comprehension sessions OR the TA did not complete REACH training.

- Second, using the predicted optimal compliance variable; 1=pupils who attended 89%+ Reading and 89%+ Comprehension sessions led by a TA who completed the REACH training, 4+ online seminars and gaps tasks; 0=control pupils plus pupils in intervention schools who attended <89% Reading and <89% Comprehension sessions OR the TA did not complete training / attend online seminars / complete gap tasks.

Please note that the specified measure of compliance is at the TA level but the final / overall measure of compliance is at the pupil level. Therefore the same approach for obtaining the CACE estimate for the specified 4-level MSCRT will be used if the TA level is not viable and the design becomes a 3-level MSCRT.

Intra-cluster correlations (ICCs)

The pre-test for REACH Primary will be KS1 Reading attainment (Y2, age 6/7) and the post-test will be the WIAT III (Y3, age 7/8). For both pre and post-test, ICCs at the school and TA levels will be estimated using a null (empty) 3-level multilevel variance components model. Within the analyses, a table will present the variance decomposition for the three levels (school, TA and pupil) along with the ICC estimates.

$$ICC_{School} = \frac{\sigma^2_{school}}{(\sigma^2_{school} + \sigma^2_{TA} + \sigma^2_{y})}; \qquad ICC_{TA} = \frac{\sigma^2_{TA}}{(\sigma^2_{school} + \sigma^2_{TA} + \sigma^2_{y})}$$

Effect size calculation

The causal impact of REACH Primary on pupil reading WIAT III will be measured using the Hedges g effect size statistic. Hedges g standardises the difference between the attainment of pupils in treatment schools and pupils in control schools into units of standard deviations. As specified in the EEF analyses guidance, the unconditional variance will be used to obtain the standard deviation. Specifically, the variance in the WIAT III outcome that is clustered at school, TA and pupil levels will be used:

$$ES = \frac{(T-C)_{adjusted}}{\sqrt{\delta^2_{sch} + \delta^2_{TA} + \delta^2_{pup}}}$$

Where:

δ^2_{sch} is the school level variance, δ^2_{TA} is the TA level variance and δ^2_{pup} is the pupil level variance for the WIAT III outcome from the empty/null multilevel model.

$(\mathrm{T}-\mathrm{C})_{\mathrm{adjusted}}$ is the mean difference between the attainment of pupils in treatment schools and pupils in control schools in the original raw WIAT III units. This is obtained from the coefficient for the school level 'group' variable from the final (headline) analyses.[4]

[4] From the model specification above this difference is estimated at the '0' of the centred independent variables on all levels. Zero at pupil level would be a pupil who attainment is the same as the mean score amongst all working with their TA; 0 at TA level would be a group of pupils under a TA that have a mean attainment the same as their school; 0 at the school level would mean a school with mean attainment (i.e. their attainment is the same as the unweighted school mean).

The coefficient standard error and the upper/lower 95% confidence intervals will also be converted into units of standard deviations using the above formula.

Appendix I - MDES calculation

From Kelcey et al (2017), the Minimum Detectable Effect Size (MDES) for a 3-level CRT is

$$MDES_{3LCRT} \sim M_{K-L-2}\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{ICC_{sch}(1-R^2_{sch})}{K}+\frac{ICC_{TA}(1-R^2_{TA})}{JK}+\frac{(1-ICC_{sch}-ICC_{TA})(1-R^2_{pup})}{nJK}}$$

This 3-level CRT equation captures much of the design but does not acknowledge that the randomisation (and training) was blocked by geographical area. Spybrook et al., (2016), call a 3-level clustered trial that is blocked by a fourth variable (geography) a 4-level Multi-site Clustered RCT (MSCRT).

Note that this design includes the geography blocking variable as a fixed effect and so the resulting multilevel model will still have three levels of random effects (School, TA and pupil).

Spybrook et al (2016) specify the MDES estimate for a 4-level MSCRT (i.e. 3-level CRT blocked by geographical area) as shown below. This assumes zero effect size variability across clusters and includes covariate explanatory power at TA and pupil levels:

$$MDES_{4LMSCRT} \sim M_{(M(K-L-2))}\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{ICC_{sch}\left(1-R^2_{sch}\right)}{MK}+\frac{ICC_{TA}\left(1-R^2_{TA}\right)}{MKJ}+\frac{(1-ICC_{sch}-ICC_{TA})\left(1-R^2_{pup}\right)}{MKJn}}$$

It can be useful to re-organise this equation following Hedges & Rhoads (2010)…

$$MDES_{4LMSCRT} \sim M_{(M(K-L-2)}\sqrt{\frac{1}{P(1-P)MKJn}}\sqrt{1+(Jn-1)ICC_{sch}+(n-1)ICC_{TA}-\left[R^2_{pup}+\left(JnR^2_{sch}-R^2_{pup}\right)ICC_{Sch}+\left(nR^2_{TA}-R^2_{pup}\right)ICC_{TA}\right]}$$

Where…

- P is the proportion of schools who receive the intervention (=0.5)
- Explanatory power at three levels is included. For the pupil level, we draw on the EEF interim test database for the estimated correlation of 0.74 (i.e. pupil level explanatory power = R^2_{pup} =0.55). For explanatory power at the TA and school levels and have estimated a lower correlation of 0.60 and corresponding explanatory power $R^2_{Sch} = R^2_{TA} = 0.36$
- ICC_{Sch} is the school level Intra Cluster Correlation coefficient (=0.14) taken from the EEF interim test database.
- ICC_{TA} is the TA level Intra Cluster Correlation coefficient (=0.02). Demack (2019) recommends a class level ICC value of 0.10 at KS2 but we have opted for a much smaller clustering at TA level to reflect how the pupils TAs will work with will be more homogenous groupings (pupils identified as struggling to read).
- M is the number of geographical sites (=5)
- K is the number of schools per site (=16)
- J is the number of TAs per school (=2)
- n is the number of pupils per TA (=5)
- L is the number of school level covariates (=6)
- $M_{(M(K-L-2))}$ is the t-distribution multiplier with M(K-L-2) (40) degrees of freedom. Assuming a two-tailed test with a statistical significance of 0.05 (α/2=0.025) and statistical power of (1-β=0.80). M_{54} = 2.8718.

This results in an MDES estimate of 0.24 standard deviations. For the FSM analyses, the number of pupils per TA is reduced to two and if all other factors are assumed to be the same as above, the FSM MDES estimate is 0.29 standard deviations.

NOTE:

These estimates assume a 3-level CRT design that is blocked by geographical area (i.e. a 4-level MSCRT as defined by Spybrook et al., 2016). The viability of the TA level is unknown at this point in time. If it is found that the TA level is not viable, the multilevel design will become a 2-level CRT blocked by geographical area (i.e. a 3-level MSCRT)

Assuming 10 pupils per school (4 FSM pupils per school), a 3-level MSCRT design results in an MDES estimate of 0.24 standard deviations overall and 0.28 sds for the FSM subsample.

Appendix II - Analysis

Overview

This Appendix provides additional details for the planned ITT analyses of the primary (headline) outcome (WIAT III) for the REACH Primary efficacy trial. Specifically, this Appendix includes:

- Specification of the multilevel regression model
- Example STATA code that will be used to fit the multilevel model

Specifying the multilevel analyses

As shown in Appendix I, the REACH Primary efficacy trial has a 3-level CRT research design blocked across (five) geographical hub areas. In addition to geographical hub area, school level KS1 Reading was used to stratify the sample. The trial design assumed that this pre-test covariate would be included at all three levels (pupil, TA and school).

To avoid multicollinearity between the three KS1 Reading covariates, they will all be centred as outlined by Hedges and Hedberg (2013). Specifically, this means that:

- Pupil level KS1 Reading will be centred around the TA level mean KS1 Reading.
- TA level KS1 Reading will be centred around the school level mean KS1 Reading.
- School level KS1 Reading will be centred around the overall (unweighted) school level grand mean[5].

This approach ensures that zero variance in the outcome will be shared across the three variables (i.e. the correlation between them will be zero).

To reflect the research design, a 3-level multilevel regression model will be fitted to the data that will aim to account for variation in the WIAT III) primary outcome. This model will include covariates at all three levels. Most covariates will be included at the school level (Intervention/Control identifier; three stratification variables & school level KS1 Reading) but KS1 Reading will also be included at both TA and pupil levels.

This will be a random intercepts model. This means that the analyses will assume that the impact of REACH Primary will be consistent across schools and TAs. As this is an efficacy trial, this assumption is appropriate (Spybrook, 2016). If this efficacy trial finds evidence of positive impact for REACH Primary on pupil reading, a future larger scale effectiveness trial may be funded that could reliably examine variation in impact across schools and TAs using multilevel models with both random intercepts and slopes.

To formally specify the ITT model, let Y_{ijk} represent the score in the WIAT III primary outcome in summer 2020 for pupil i for TA j in school k.

The level 1 (pupil level) model is:

Equation II.1 $$Y_{ijk} = \pi_{0jk} + \pi_{1jk}\left(KS1\_pc_{ijk}\right) + e_{ijk} \qquad e_{ijk} \sim N(0, \sigma^2)$$

Where:

- $i = 1, \ldots, n$ pupils per TA; $j = 1, \ldots, J$ TAs per school; $k = 1, \ldots, K$ schools
- π_{0jk} is the mean score for TA j in school k
- $KS1\_pc_{ijk}$ is the pupil level (TA-centred) KS1 Reading pre-test covariate for pupil i in TA j in school k. π_{1jk} is the coefficient for the pupil level KS1 Reading covariate for TA j in school k

[5] The unweighted school level grand mean is the mean obtained using all school means. This means that each school mean will count once in calculating the unweighted school level grand mean. An overall pupil-level mean would be weighted at the school level by the number of pupils in each school.

- e_{ijk} is the pupil level error/residual
- σ^2 is the residual/error variance between pupils within-TA

The level 2 (TA level) model is:

Equation II.2 $$\pi_{0jk} = \beta_{00k} + \beta_{01k}\left(KS1\_clc_{0jk}\right) + r_{0jk} \qquad r_{ijk} \sim N(0, \tau_\pi)$$

Where:

- β_{00k} is the mean score for school k
- $KS1\_clc_{0jk}$ is the mean TA level KS1 Reading covariate (school centred) for TA j in school k. β_{01k} is the coefficient for the TA level KS2 covariate for school k
- r_{0jk} is the random effect associated with each TA
- τ_π is the residual/error variance between TAs within schools

The level 3 (School level) model is:

Equation II.3 $\beta_{00k} = \gamma_{000} + \gamma_{001}Group_k + \gamma_{002}KS1\_Sch_k + \gamma_{003}Hub_k + u_{00k};\ u_{00k} \sim N(0, \tau_\beta)$

Where:

- γ_{000} is the estimated adjusted grand mean
- $Group_k$ is '1'for treatment and '0' for control schools, γ_{001} is the effect of REACH Primary participation
- $KS1\_Sch_k$ is the school level mean KS1 Reading covariate (centred around the school level mean), γ_{002} is the coefficient for the school level KS1 Reading covariate.
- Hub_k represents the stratification variable (geographical hub area,), γ_{003} is a coefficient vector for the school level stratification covariates (geographical hub as dummies). In total, for the stratification variables, four binary dummy variables will be included (four to account for the five geographical hub areas)
- u_{00k} is the random effect associated with each school mean
- τ_β is the residual/error variance between schools

Example of STATA SYNTAX that will be used to fit the multilevel model

The multilevel regression model will be fitted to the data using the Stata mixed command, an example of the code that will be used is shown below:

Empty / Null Model:

° mixed WIAT || School_ID: || TA_ID:

Outcome Only:

° mixed WIAT Group || School_ID: || TA_ID:

KS1 to WIAT Progress:

° mixed WIAT Group KS1Reading_SchC KS1Reading_TAC KS1Reading_PupC || School_ID: || TA_ID:

Final (headline) analyses:

° mixed WIAT Group KS1Reading_SchC KS1Reading_TAC KS1Reading_PupC b1.Hub || School_ID: || TA_ID:

The empty/null model will be used to obtain the standard deviation for calculating the Hedges g effect size statistic.

References

Gerber, A., & Green, D. (2012). Field Experiments: Design, analysis and Interpretation. W.W. Norton & Company.

Hedges, L.V. & Hedberg, E.C. (2013) Intraclass correlations and covariate outcome correlations for planning two- and three-level cluster randomised experiments in education. Evaluation Review 37(6) pp445-489.

Hedges, L.V. & Rhoads, C. (2010) Statistical power analysis in education research. NCSER 2010-3006. Available at https://ies.ed.gov/ncser/pubs/20103006/pdf/20103006.pdf

Kelcey, B., Spybrook, J., Phelps, G., Jones, N. & Zhang, J. (2017) Designing large scale multisite and cluster randomized studies of professional development. The Journal of Experimental Education, 85(3) pp389-410

Spybrook, J., Shi, R., Kelcey, B. (2016) Progress in the past decade: an examination of the precision of cluster randomized trials funded by the U.S. Institute of Education Sciences. International Journal of Research and Method in Education. 39 (3) pp255-267