Infant Language Link
Statistical Analysis Plan

Evaluator (institution): Sheffield Hallam University
Principal investigator(s): Dr Martin Culliney

Education Endowment Foundation

Template last updated: August 2019

PROJECT TITLE	Evaluation of Infant Language Link, a two-armed cluster randomised trial
DEVELOPER (INSTITUTION)	Speech Link Multimedia Ltd.
EVALUATOR (INSTITUTION)	Sheffield Hallam University
PRINCIPAL INVESTIGATOR(S)	Dr Martin Culliney
PROTOCOL AUTHOR(S)	Dr Martin Culliney, Dr Ester Ehiyazaryan-White, Dr Dieuwerke Rutgers
TRIAL DESIGN	Two-arm cluster randomised controlled trial with random allocation at school level
TRIAL TYPE	Efficacy
PUPIL AGE RANGE AND KEY STAGE	Age 5-6, KS1
NUMBER OF SCHOOLS	166
NUMBER OF PUPILS	3021
PRIMARY OUTCOME MEASURE AND SOURCE	Language and communication Pearson CELF-5 (sentence comprehension and linguistic concepts subtests) and Renfrew Action Picture Test - combined
SECONDARY OUTCOME MEASURE AND SOURCE	Language and communication Pearson CELF-5 (sentence comprehension and linguistic concepts subtests) and Renfrew Action Picture Test - separate

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0 *[original]*	08/05/2024	*N/A*

Table of contents

Introduction

Infant Language Link enables schools to identify and support children in reception, Y1 and Y2 with mild to moderate language needs. The intervention uses a tiered structure that incorporates both whole class provision and separate additional support for pupils found to be in need, who are identified through a standardised universal screening tool that assesses receptive language. Teachers deliver the universal element of the intervention to all pupils in class. Targeted group interventions are also delivered to groups of 4-5 pupils by TAs. The amount and type of intervention depends on the child's performance on the initial screening. Further information about the intervention can be found in the evaluation protocol.

Any school with at least 20 pupils in the 2023/24 Y1 cohort was eligible to join the trial, excluding those already using the intervention, which is available commercially. A maximum of 20 pupils per school are included in the evaluation. Schools supplied details on all pupils in one Y1 class to the evaluation team who then randomly selected 20 pupils to participate in baseline assessments. There was no guidance to schools on which of their Y1 classes to select with regard to ability or any other factors; schools were allowed to choose. This was to ensure that the class teacher was willing to take part. Assessing more pupils per school would have made little difference in statistical sensitivity and the additional costs would have been difficult to justify. The whole class provision will still be delivered to the whole class, but only the sampled pupils will participate in the evaluation data collection and assessment.

Pupil data was provided prior to randomisation, which took place on 3 November 2023. Baseline assessments were completed in all schools by this date, except two which were postponed at short notice due to assessor absence. Schools learned their allocation once baseline data collection was completed. This notification was delayed until 8 November 2023 for the two schools whose final assessments were postponed until this date.

The headline analysis sample for this trial will be the whole sample of pupils randomly selected by the evaluators, with a maximum of 20 per school. Pupils in intervention schools identified for further support will be treated as a subgroup. This is important to measure the effects of both the targeted and whole class components of the intervention.

The primary outcome will be a combined measure of language ability, constructed from the pupil assessments conducted in schools. As the intervention targets receptive and expressive language but does not prioritise one or the other it was agreed that the primary outcome should incorporate both. In the absence of an assessment that measures receptive and expressive language together, a combined measure derived from separate assessments was deemed most suitable, following precedent from previous evaluations. Full details of all outcome measures and subgroup analyses are provided below.

Design overview

Table 1: Trial design

Trial design, including number of arms		Two-arm, cluster randomised
Unit of randomisation		School
Stratification variables (if applicable)		Education Investment Area status (Yes/No) Use of any relevant interventions (Yes/No) Use of external speech and language therapy support (None/low frequency/half-termly or more frequent)
Primary outcome	variable	Language and communication combined measure
	measure (instrument, scale, source)	Pearson CELF-5 (sentence comprehension and linguistic concepts subtests) and Renfrew Action Picture Test (Information and Grammar) combined
Secondary outcome(s)	variable(s)	Language and communication separate measures
	measure(s) (instrument, scale, source)	Pearson CELF-5 (sentence comprehension 0-26) Pearson CELF-5 (linguistic concepts 0-25) Renfrew Action Picture Test (Information 0-41, grammar 0-39)
Baseline for primary outcome	**variable**	Language and communication
	measure (instrument, scale, source)	Pearson CELF-5 (sentence comprehension and linguistic concepts subtests) and Renfrew Action Picture Test combined
Baseline for secondary outcome	**variable**	Same as secondary outcomes listed above

Sample size calculations overview

The design is a 2-level clustered RCT. In calculating the Minimum Detectable Effect Size (MDES), the smallest effect size that could be detected as statistically significant (often set as p<0.05) with a statistical power of 80% or higher, our estimates at the protocol stage were based on the following assumptions:

M_{j-k-2} - T-distribution multiplier assuming a two-tailed test with a statistical significance of 0.05, statistical power of =0.80 and J-K-2 (164) degrees of freedom
R_i - Participant (pupil) level pre/post-test correlation of 0.75 (R_i^2 =0.56)
R_c - Cluster (school) level pre/post-test correlation of 0.20 (R_c^2 = 0.04)
ρ - Intracluster correlation (ICC) of 0.20
j - Number of schools = 170
m - Pupils per school = 20
k - Number of cluster level covariates[1] = 4
P - Proportion of schools allocated to intervention group (P=0.5)

The participant correlation estimates in the protocol were taken from the most recent EEF evaluation of NELI (Dimova et al 2020), which used similar primary outcome measures (Preschool CELF instead of CELF-5, along with the Renfrew Action Picture Test). The school level correlation was conservatively estimated as 0.20. As the ICC reported at the analysis stage of Dimova et al (2020) was surprisingly high (0.35), a lower ICC was assumed (0.20). This is closer to the figures from the randomisation and protocol stages of Dimova et al, which were 0.15 and 0.12 respectively. An ICC of 0.20 is also the default value recommended for attainment outcomes by the IES What Works Clearinghouse (2022:171). Since the power calculations were published in the protocol, baseline data has become available, and the unconditional ICC is 0.14. This suggests that the ICC estimate from the protocol may have been overcautious.

Calculations were performed in Excel using the formula set out in Bloom et al (2007) for two-level clustered randomised controlled trials and checked using powerup! (Dong and Maynard 2013). This allows covariates to be included at both individual (pupil) and cluster (school) level, which in turn increases sensitivity. The MDES equation is:

$$\mathrm{MDES} = M_{j-k-2}\sqrt{\left(\frac{\rho(1-R_c^2)}{\mathrm{P(1-P)J}}\right)+\left(\frac{(1-\rho)(1-R_i^2)}{\mathrm{P(1-P)Jm}}\right)}$$

Table 2 (below) summarises the MDES estimates for the central design based upon the estimates and assumptions outlined above for a sample with 170 schools and 20 pupils per school to reflect the recruitment target and the minimum class size eligible for the trial. These figures were used in the protocol, where the MDES estimate is 0.20. This is similar to the last NELI evaluation, which reported an MDES of 0.19 (Dimova et al 2020:19). For the FSM subgroup, estimated at five pupils per school, the MDES is 0.22 for 170 schools.

At randomisation, there were 166 schools with an average of 18 pupils per school. Updating the power calculations with this sample size also produces an MDES of 0.20. As pupil FSM status is being obtained from the NPD, at the time of writing it is only possible to estimate the

[1] Whether a school is in an Education Investment Area, uses another relevant intervention, uses external speech and language support at least once each half term, or uses such support but less frequently than every half term.

number of FSM pupils in each school. Assuming that there are five FSM pupils in the analysis sample of pupils from each school, the trial is powered to MDES 0.22 for this subgroup, rising to 0.23 for four FSM pupils, and 0.24 for three pupils. The power calculations for FSM pupils were conducted without prior expectations owing to the general lack of evidence for Y1 pupils and the absence of FSM analysis in most recent NELI report which was the most similar evaluation found when designing the trial.

To illustrate the robustness of this design to the impact of attrition, indicative MDES estimates are provided. MDES estimates assume that randomisation has been maintained, an assumption undermined by attrition. While these indicative MDES estimates are useful for illustrating the robustness of RCT design, they need to be treated cautiously because they assume that any attrition will be random. With 10% attrition at school level (leaving 149 schools), the indicative MDES is 0.21. This would remain applicable provided that pupil level attrition stays below 20%. In other words, with at least 149 schools and 15 pupils from each school in the analysis sample result in an indicative MDES of 0.21. With no attrition at school level, the MDES would remain at 0.20 if at least 13 pupils per school are present in the analysis sample.

Table 2: Sample size calculations

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
Minimum Detectable Effect Size (MDES)		0.20	0.22	0.20	0.22
Pre-test/ post-test correlations	level 1 (pupil)	0.75	0.75	0.75	0.75
	level 2 (school)	0.20	0.20	0.20	0.20
Intracluster correlations (ICCs)	level 2 (school)	0.20	0.20	0.20	0.20
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		2	2	2	2
Average cluster size		20	5	18	5
Number of schools	intervention	85	85	83	83
	control	85	85	83	83
	total	170	170	166	166
Number of pupils	intervention	1700	425	1500	415
	control	1700	425	1521	415
	total	3400	850	3021	830

Analysis

Multilevel linear random intercept models will be constructed for the primary outcome, with pupils clustered into schools, using the 'mixed' command in Stata (versions 15 onward). A measure of language and communication ability combining the selected Pearson CELF-5

subtests (sentence comprehension and linguistic concepts) and the Renfrew Action Picture Test (henceforth RAPT) will be used as the baseline covariate for analyses of the primary outcome. This will be the same as the outcome measure but collected at pre-intervention.

A complete cases approach will be used in all analyses. The first model will only include the school level group identifier (an outcome only model) and will supply the unconditional variance figure used to calculate the effect size. The second model will add the baseline covariate at the pupil and school levels[2]. The final model will also include the randomisation stratifiers: whether the school is in an Education Investment Area, use of similar interventions, use of external speech and language therapy services. This model will form the headline ITT impact analysis for the primary outcome. Results of all three models for the primary outcome will be presented in an appendix to the report so the difference between the effect sizes can be compared, however for all other analyses only the results from the final model will be included in the main report. The headline effect size will be calculated using the coefficient from the following equation:

$$Y_{ij} = b_0 + b_1 Group_j + b_2 Baseline_{ij} + b_3 Baseline\_school_j + b_4 EIA_j + b_5 Uses\_int_j + b_6 SLT\_low_j + b_7 SLT\_high_j + u_j + e_{ij}$$

Where Y_{ij} is the outcome for pupil *i* in school *j*, b_0 is the constant, and *Group* is a binary indicator of school treatment allocation. Pupil and school level baseline covariates are represented by *Baseline* and *Baseline_school*. The stratifiers used in the randomisation are denoted as *EIA,* a binary indicator of whether the school is in an Education Investment Area, *Uses_int,* a binary indicator of whether the school uses a relevant intervention apart from Infant Language Link, and *SLT_low* and *SLT_high*, dummies derived from the categorical variable showing whether schools use external speech and language therapy support. The random intercepts are represented by u_j, and e_{ij} is the error term.

For each model, the coefficient of the school-level dummy variable used to distinguish 'intervention group' pupils within the schools who will receive the Infant Language Link programme from 'control group' pupils will be converted into Hedges' g effect size statistics with 95% confidence intervals.

Primary outcome analysis

The intervention aims to improve pupil expressive and receptive language. Due to the equal importance of these two dimensions, it was agreed during the evaluation setup period that the primary outcome should combine both. Specifically, the measure will comprise the two Pearson CELF-5 subtests that are most relevant to this intervention (sentence comprehension, scored on a 0-26 scale, and linguistic concepts, scored 0-25), and the Renfrew Action Picture Test, which is scored in two parts (information, scored 0-41, and grammar, scored 0-39).

The evaluation protocol stated that the intention is to combine these measures into a latent language variable using structural equation modelling, as per the approach used in other recent evaluations in this area (Dimova et al, 2020; Menzies et al, 2022). However, attempting this with baseline data collected for this trial showed a poor fit with the data. The

[2] These will be centred so that the school level mean will be centred on the mean for all schools and the pupil level will be centred around the school mean (see Hedges and Hedberg 2013).

goodness of fit indicators did not meet the thresholds recommended in Hu and Bentler (1999). Further detail is provided in the Appendix.

It was decided that the approach used in the original EEF NELI trial (Sibieta et al 2016), which combined language measures by standardising and summing the constituent scales and was intended as a sensitivity analysis for this trial, should be used to create the primary outcome. Specifically, each of the four language scales will be converted into a z score, these will all be added and the resulting value converted into a z score. The process of constructing the primary outcome measure is detailed in the Appendix along with the Stata code.

All outcome data will be collected during June/July 2024. Assessments will be administered in school by speech and language therapists. The CELF-5 subtests are completed first, followed by the RAPT. The developer advises that conducting a receptive language assessment before an expressive language assessment makes sense as the latter requires children to talk and can be seen as more demanding. Duration varies between pupils but is not expected to exceed 30 minutes in total per child. All test administrators are blinded to allocation, although it is possible that some schools will inadvertently reveal their allocation in the course of the assessment visit.

Assessors are required to attend a half-day training delivered by the evaluator, which consists of practical demonstration and role-play practice of the CELF-5 and RAPT. This is expected to improve consistency between the different testers, who will carry out data collection in person at participating schools and record their marks electronically before posting all completed test papers back to the evaluation team. Testers are being recruited on the basis of their professional qualifications and experience in administering the selected assessments. A team of 38 assessors completed the baseline assessments and 35 have been recruited for the outcome testing, 21 of whom were involved at baseline. Quality assurance processes to ensure reliability will include moderation of a sample of 5% of returned test papers.

For each analysis, the assessment data collected post-intervention will be used as the outcome. The same measure collected at pre-intervention will be used as the baseline covariate. All baseline assessments were conducted in schools between 18 September and 8 November 2023.

Secondary outcome analysis

The two CELF-5 subtests used for the primary outcome will be analysed separately as secondary outcomes. The Renfrew Action Picture Test also comprises two components (information, scored 0-41, and grammar, scored 0-39) which will be analysed separately as secondary outcome measures. For each secondary outcome the raw score will be used. While the developers stress that the intervention is equally relevant to expressive and receptive language, these analyses will highlight any evidence of variation in effects for these different areas of pupil development.

Subgroup analyses

Subgroup analysis will be conducted on the following restricted samples, using the approach outlined above for the headline primary outcome:

- Pupils selected for additional small group support in intervention schools. These pupils are identified by the initial screening as having mild to moderate speech and language needs. A comparison sample from control schools will be identified using scores on the baseline assessment. Correlations between screening assessment scores and baseline assessment scores will be presented for the intervention group to provide assurance that this approach selects an appropriate comparison sample.
- Pupils eligible for free school meals as identified by the 'EVERFSM_6' indicator obtained from the NPD, as is recommended for all EEF trials.
- Pupils with English as an additional language as defined in the NPD, as the developer provides specific advice for supporting these pupils.

The developer recommends that fluency in English is recorded for EAL pupils. We will request this data and explore options for sensitivity analysis, such as using the fluency level as an additional covariate in EAL subgroup analysis should this be available for a majority of pupils. These analyses are exploratory as the trial is not powered for them. However, we are interested in whether the effects are different for these pupils as the intervention is focussed on language.

As per EEF guidance, FSM status will also be specified as an interaction term in an additional model and the results presented in an appendix of the final report. This model will contain the full set of covariates included in the full analysis model described above.

Additional analyses

An outcome only model and a model with only school and pupil level baseline covariates will be estimated for the primary outcome as robustness checks. Results will be presented in the appendix of the final report. The sensitivity analysis using a combined outcome measure derived from SEM will also be presented in the report.

Longitudinal follow-up analyses

None planned.

Imbalance at baseline

Table 3 shows the balance between the intervention and control groups on selected key variables at baseline. The three stratifiers used in the randomisation process are presented first. Two-thirds of schools in the sample are located in Education Investment Areas. These are evenly distributed by treatment allocation. A similar proportion of schools reported that they use at least one of the other language and communication interventions[3] currently available for the relevant age group, although this was slightly higher among control schools (69%) than intervention schools (66%). The reported frequency of schools using external speech and language therapy support was also well balanced, with 55% of intervention schools using this at least every half term compared to 54% of control schools.

School Ofsted rating was not used as a stratifier and appears to be less well balanced. This is perhaps due to the confounding effect of missing data on this variable, with more control schools (n=9) than intervention schools (n=3) having a missing value. The number of schools rated as good are almost equal between the two treatment conditions and while

[3] These interventions were NELI, Talk Boost, Wellcomm, Language for Learning, and Elklan.

there is greater imbalance in the number of schools rated as outstanding or requiring improvement, the numbers in these categories are relatively small.

Table 3: Imbalance at baseline

	Baseline (N Schools=166)		Analysis (N Schools=)	
	Intervention (N=83)	Control (N=83)	Intervention (N=)	Control (N=)
School level (categorical)	% (n)	% (n)	% (n)	% (n)
Stratifiers				
Not EIA	35%(29)	34%(28)		
EIA	65%(54)	66%(55)		
Does not use interventions	34%(28)	31%(26)		
Uses relevant interventions	66%(55)	69%(57)		
No external SLT	10%(8)	11%(9)		
Infrequent external SLT	35%(29)	35%(29)		
External SLT at least half-termly	55%(46)	54%(45)		
School type				
Academies	39%(32)	34%(28)		
Free schools	0%(0)	6%(5)		
Local authority	61%(51)	60%(50)		
Urban/ rural status				
Rural	13%(11)	11%(9)		
Urban	87%(72)	89%(74)		
OFSTED ratings				
Outstanding	10%(8)	7%(6)		
Good	76%(63)	76%(63)		
Requires Improvement	11%(9)	6%(5)		
Missing	4%(3)	11%(9)		
School level (continuous)	Mean (SD)	Mean (SD)		
Total number of pupils (including part-time pupils)	363(209)	290(160)		
Percentage of disadvantaged pupils	33.71(14.1)	35.15(14.3)		
KS1 average points	7.59(0.42)	7.52(0.44)		
Pupil level (continuous)	Mean (SD)	Mean (SD)	Effect size*	
Pre-test scores				
CELF LC	16.01(5.59)	16.44(5.32)	-0.03	
CELF SC	15.92(6.16)	16.43(5.77)	-0.03	
RAPT Grammar	21.83(6.68)	22.08(6.5)	-0.01	
RAPT Information	26.76(6.2)	27.56(5.94)	-0.03	
Combined language (SEM)	-0.16(3.40)	0.16(3.14)	-0.10	
Combined language (z scores)	-0.05(1.05)	0.05(0.94)	-0.10	
Combined language (alpha)	-0.04(0.87)	0.04(0.78)	-0.10	

Another area with imbalance between the intervention and control groups is the number of pupils per school. The three all through schools in the sample distort the overall figures somewhat, although removing them would still leave the average size of intervention schools noticeably larger than control schools, so they are included in the figures presented.

However, the percentage of FSM pupils is well balanced across both treatment allocations, perhaps reflecting the use of Education Investment Area status as a stratifier.

Finally, looking at the four outcome assessment measures separately shows that each of them is imbalanced at baseline, with higher scores for the control group. It is therefore unsurprising that for each of the three methods to combine the four scales into a single language variable (see Appendix for details), scores are higher for the control group. These methods all produce similar results. As randomisation took place after the completion of baseline assessments, the difference in scores between the intervention and control groups has probably emerged by chance.

Missing data

There were no missing outcome data at baseline as only pupils that completed the assessments are included in the study sample. The only possible missing data will therefore be found in the outcomes at post-intervention. The reasons for any missing data (such as school/pupil withdrawal) will be summarised in the final report.

The impact analyses will examine missing data in the outcome and explanatory variables and consider whether it is reasonable to assume that the missing data are random. A multilevel logistic regression model with a binary outcome identifying when outcome data is missing (=1) or not (=0) and the same covariates as the headline ITT model will be estimated to examine any patterns. This model will then be replicated with only participants at schools that took part in the outcome testing, to focus on pupil level attrition.

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. If over 5% of pupil outcome data is missing, 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 will be used to identify whether the missing outcome data can be assumed to be missing completely 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 completely at random, otherwise multiple imputation will be used and the results compared with the headline ITT analysis for the primary outcome.

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

Compliance will be measured at the school level. The three indicators are attendance at training, administering initial language screening to pupils, and delivering targeted group sessions. On the first two, compliance is expected to be near to 100% as both are scheduled at the very start of delivery period and can be completed quickly.

The listening group works on developing the underpinning attention and listening skills required to support children's language development. Schools are asked to complete a

minimum of six sessions of this group, and after this time if the children are fully achieving the activities and overall aims for the group, they can choose to finish the group early. In contrast the language groups are working on a variety of different language targets and all of the sessions need to be completed in order for the children to make good progress against the aims of the group. For this reason, all eight sessions in the language group must be delivered for a school to achieve compliance.

The group sessions are delivered throughout the school year and demonstrate that the school is implementing the programme after the training and initial pupil screening. Details of the number of sessions required for a school to achieve full or part compliance can be found in Table 4 (below). If no participating pupils in a given school are identified as requiring further support, the school will not be included in the compliance analysis.

These variables will be used to estimate the Complier Average Causal Effect (CACE). The purpose of the CACE analysis is to estimate the impact of Infant Language Link for pupils in schools deemed to have 'complied' with the intervention. CACE will be estimated using two-stage least squares (2SLS) regression (Gerber and Green, 2012). The first stage will model compliance using the randomisation stratifiers along with additional school level items that are available via the school census as listed above in Table 3. This will be a multilevel logistic regression model used to generate predicted 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. This process will be undertaken twice, for part and full compliance.

Table 4: Compliance indicators

Activity	Full compliance	Part compliance
Training	SENCo attends 3 initial sessions; teacher and TA attend 2 sessions each	
Initial pupil language screening	Delivered to all pupils in participating class at start of intervention period	
Targeted groups	Delivering 6/8 sessions in a listening group, and 8/8 sessions in two others	6/8 in listening group plus one other group

Intra-cluster correlations (ICCs)

For the primary outcome at both pre- and post-intervention, ICCs at the school level will be estimated using a null (empty) 2-level multilevel variance components model. Within the analyses, a table will present the variance decomposition for the two levels (school and pupil) along with the ICC estimates. The ICC for the full headline analysis model of the primary outcome will also be presented.

$$\mathrm{ICC} = \frac{\mathrm{Variance}_{\mathrm{school}}}{\mathrm{Variance}_{\mathrm{school}} + \mathrm{Variance}_{\mathrm{pupil}}}$$

Effect size calculation

The effect size measure to be used will be Hedges' g. This will be calculated using the following equation.

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

Where:

δ_{sch}^2 is the school level variance and δ_{pup}^2 is the pupil level variance for the language 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. This is obtained from the coefficient for the school level treatment allocation variable from the final headline analyses.

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

References

Bloom, H, S., Richburg-Hayes, L. and Black, A. R. (2007) 'Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions' *Educational Evaluation and Policy Analysis* 29(1) 30–59

Dimova, S., Ilie, S., Rosa Brown, E., Broeks, M., Culora, A. and Sutherland, A. (2020) *The Nuffield Early Language Intervention: Evaluation Report.* Education Endowment Foundation

Dong, N., Kelcey, B., Maynard, R. and Spybrook, J. (2015) *PowerUp! Tool for power analysis*.

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

Hedges, L.V. and 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.

Hu L., Bentler P. M. (1999). 'Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives' *Structural Equation Modeling: A Multidisciplinary Journal*, 6, 1-55

Lai, K. and Green, S.B. (2016) 'The Problem with Having Two Watches: Assessment of Fit When RMSEA and CFI Disagree' *Multivariate Behavioral Research*, 51:2-3, 220-239

Menzies, V., Cramman, H., Eerola, P., Hugill-Jones, J., Akhter, N., and Einbeck, J. (2022) *Parents and Children Together (PACT) Evaluation Report.* Education Endowment Foundation

Sibieta, L., Kotecha, M. and Skipp, A. (2016) *Nuffield Early Language Intervention Evaluation report and executive summary.* Education Endowment Foundation

What Works Clearinghouse (2022) *What Works Clearinghouse Procedures and Standards Handbook, Version 5.0*, Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (NCEE). Available online: https://ies.ed.gov/ncee/wwc/Handbooks

Appendix A – constructing a combined language outcome

To start exploring the relationship between the separate scales derived from the pupil assessments undertaken at baseline, pairwise correlations between the variables were examined. As would be expected, the two CELF scales and the two RAPT scales correlate most strongly with one another. However, the CELF scales also correlate with the RAPT scales to at least 0.5. This suggests some relationship between receptive language, as measured by the two CELF subscales, and expressive language, as measured by the RAPT.

The four scales also have a Cronbach's alpha of 0.85. Again, this suggests that the four language scales all represent a single underlying language variable. The 'alpha' command in Stata 17 has an option to create a new variable from the items specified, using the generate subcommand. The resulting variable was also included in Table 3 above to compare with other methods of combining the four language scales into a single measure.

Table 5: Pairwise correlations between outcomes

	Celf LC	Celf SC	RAPT Grammar	RAPT Information
Celf LC	1			
Celf SC	0.65	1		
RAPT Grammar	0.52	0.50	1	
RAPT Info	0.52	0.49	0.78	1

Cronbach's alpha 0.85. N= 3021

Running a factor analysis in Stata 17 with the four language scales entered as the source variables produces the results displayed in Table 6. The eigenvalues indicate that a single factor solution would make the most sense. This is further evidence that the four language scales all represent a single underlying language variable.

Table 6: Factor analysis results

Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	2.29	2.05	1.04	1.04
Factor2	0.24	0.39	0.11	1.15
Factor3	-0.15	0.03	-0.07	1.08
Factor4	-0.18	.	-0.08	1

LR test: independent vs. saturated: chi2(6) = 5801.05 Prob>chi2 = 0.0000

As mentioned above, a single standardised scale from the four sub scales can be created using the generate option with the 'alpha' command in Stata 17. It is also possible using the approach adopted in Sibieta et al (2016), where the four scales were standardised and summed, with the resulting variable then standardised again. Our preference is to use the latter method for constructing the primary outcome measure for this trial. A key priority is to create a combined scale that gives equal weight to the different language scales. This would achieve that aim. The code for generating this outcome variable is presented in Appendix B.

A combined outcome variable derived through structural equation modelling was also considered. Fitting a structural equation model with the four language variables results in goodness of fit indicators that are short of the recommended thresholds (RMSEA below 0.06, and CFI of at least 0.95, see Hu and Bentler 1999). The basic specification with no covariances added has RMSEA 0.34 and CFI 0.88. Adding one covariance increases the CFI to above 0.9 but also increases the RMSEA to 0.42 or 0.43 depending on which

covariance is added. There is some disagreement over which cut off points to apply (see Lai and Green 2016:220 for a summary), yet these RMSEA values would not be deemed satisfactory by any measure. Table 7 below shows the goodness of fit values for the SEM with no covariances, and then for a series of models each incorporating one covariance, between one of the CELF subtests and one of the RAPT scoring components.

Further examination of the data included adding a covariance between both CELF subtests, and between both RAPT components. The results are very different, indicating what appears to be a good fit (CFI = 1, TLI = 1, RMSEA < .001). However, in this trial we would expect all pairs of language scales to correlate, but with stronger correlations between the two CELF scales and two RAPT scales. This is corroborated by the correlation coefficients presented in Table 5 above.

When adding two covariances, one for both RAPT scales and another for both CELF scales, the model will not converge. This means that one pair or the other would need to be selected but as both pairs are correlated it does not seem tenable to choose between them.

Table 7: Comparison of goodness of fit indicators for different SEM models (with and without covariances)

Fit statistic	Basic	Celf_LC* R_Info	Celf_SC* R_Info	Celf_LC* R_Grammar	Celf_SC* R_Grammar
Likelihood ratio					
chi2_ms(2)	705.87	565.37	536.35	536.35	565.37
p > chi2	0.00	0.00	0.00	0.00	0.00
chi2_bs(6)	5805.21	5805.21	5805.21	5805.21	5805.21
p > chi2	0.00	0.00	0.00	0.00	0.00
Population error					
RMSEA	0.34	0.43	0.42	0.42	0.43
90% CI, lower bound	0.32	0.40	0.39	0.39	0.40
upper bound	0.36	0.46	0.45	0.45	0.46
pclose	0.00	0.00	0.00	0.00	0.00
Information criteria					
AIC	72567.18	72428.68	72399.66	72399.66	72428.68
BIC	72639.34	72506.86	72477.84	72477.84	72506.86
Baseline comparison					
CFI	0.88	0.90	0.91	0.91	0.90
TLI	0.64	0.42	0.45	0.45	0.42
Size of residuals					
SRMR	0.09	0.07	0.07	0.07	0.07
CD	0.88	0.97	0.97	0.99	0.96

As the SEM model does not show a good fit of the data, adopting a different approach to combining the different language measures into a single primary outcome is justified. Ordinarily when SEM produces disappointing results, improved model fit could be sought by removing or adding certain items, yet here that is not an option. None of the separate language scales can be omitted as the different assessments all measure different dimensions of language and communication ability which are equally relevant to the intervention. Removing individual items from the scales is not tenable as each is part of an established measure that is not validated for such selective use.

Table 3 above shows that all combined language outcome variables are imbalanced at baseline with effect sizes of -0.10 irrespective of how the separate scales are combined. However, as all baseline assessment data was collected prior to randomisation, this is likely to be due to chance. Schools or assessment administrators were not aware of treatment allocation at the time of data collection, so it is not possible that baseline assessment performance was affected by this.

Figure 1: CELF Linguistic Concepts subtest

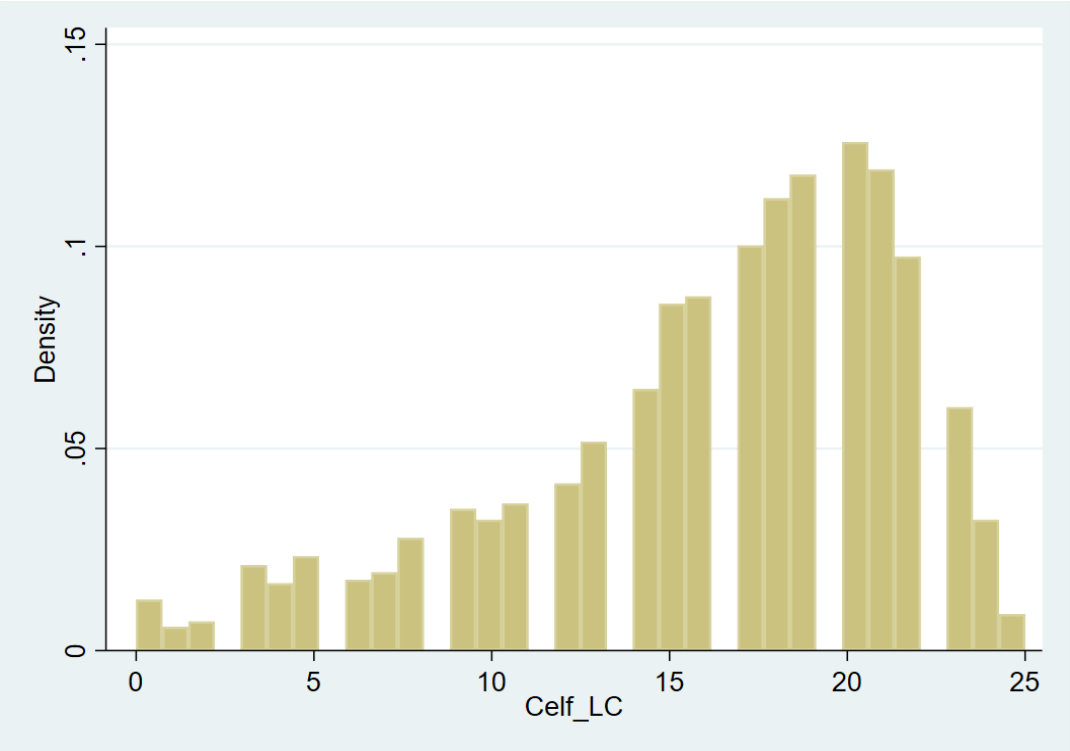


Figure 2: CELF Sentence Comprehension subtest

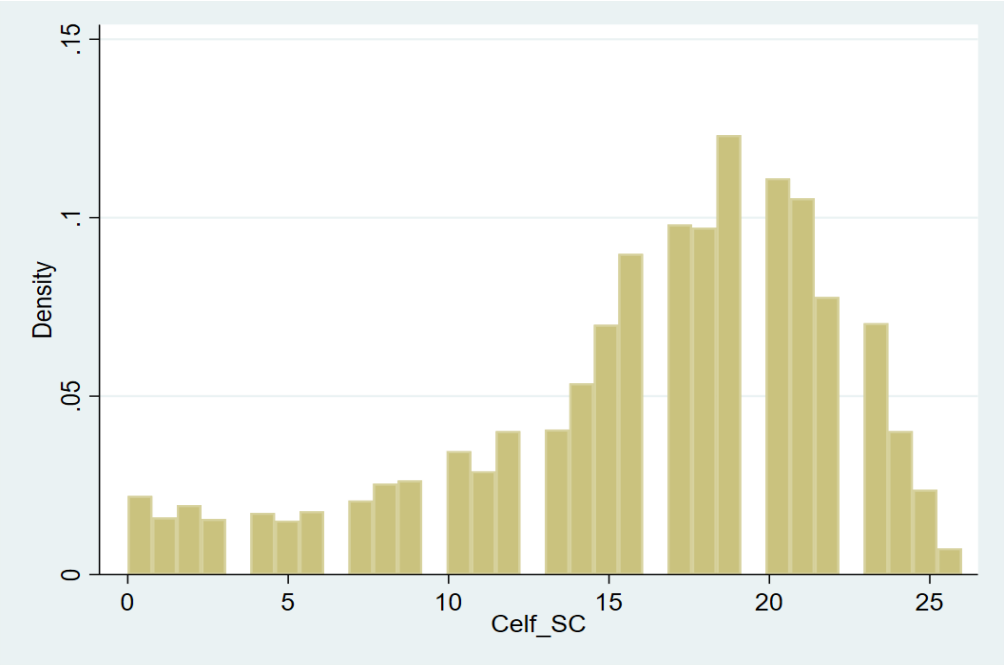


Figure 3: RAPT Grammar

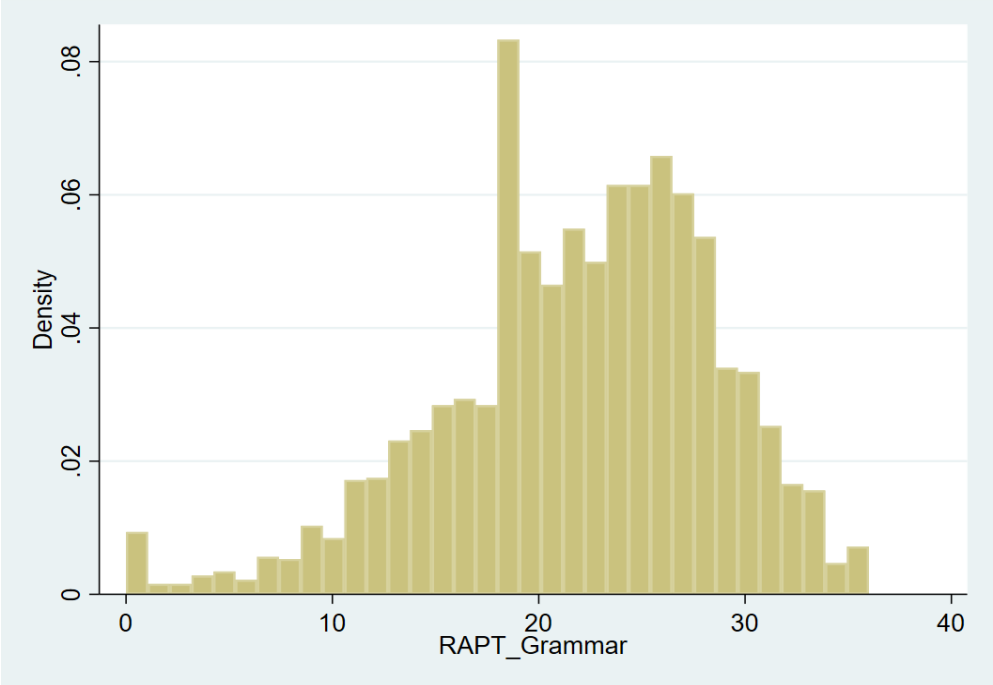
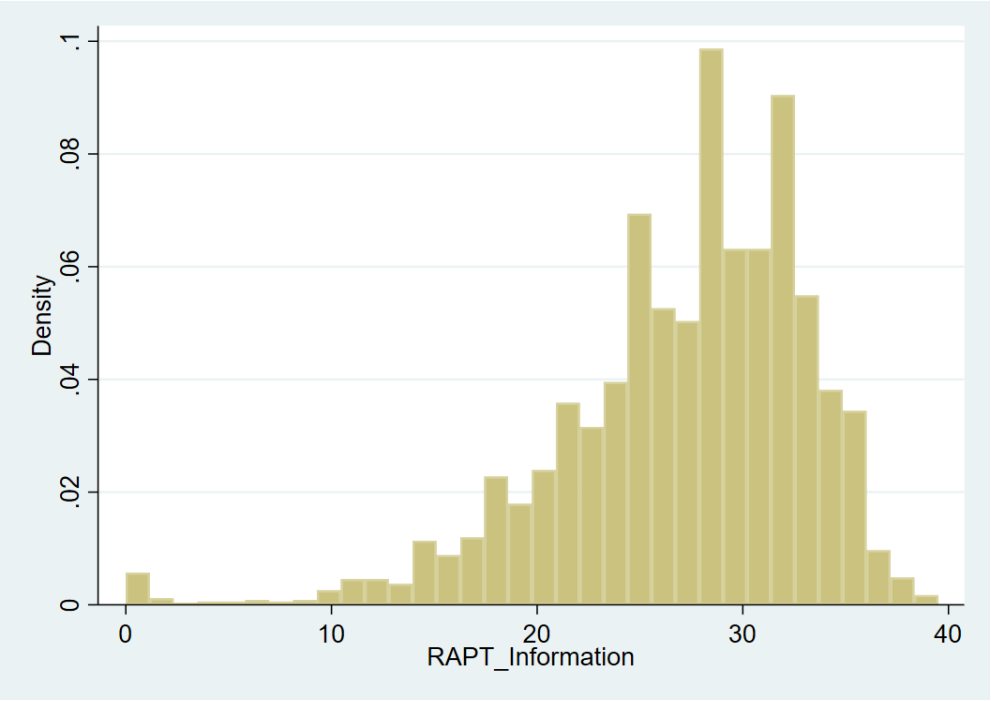


Figure 4: RAPT Information



Appendix B – example Stata code

Impact analysis

```
mixed Primary_outcome Allocation Pupil_centred_Baseline ///
centred_school_mean_Baseline  EIA Use_int EXT_SLT_low EXT_SLT_high || SchoolID:

estat icc
```

Creating combined language outcome

```
foreach x of varlist Celf_LC Celf_SC RAPT_Grammar RAPT_Information {
                                egen z_`x' = std(`x')
}

egen z_baseline = rowtotal(z_Celf_LC z_Celf_SC z_RAPT_Grammar z_RAPT_Information)

egen Z_baseline = std(z_baseline)
```

Creating combined language outcome through SEM

```
sem (Celf_LC Celf_SC RAPT_Grammar RAPT_Information <- baseline_language_latent),
nocapslatent latent(baseline_language_latent)

estat gof, stats(all)

predict baseline_language_latent, latent(baseline_language_latent)
```