Statistical Analysis Plan
[*PACT-3*]

Evaluator (institution): Durham University

Principal investigator(s): Victoria Menzies

Education Endowment Foundation

PROJECT TITLE	Evaluating the impact of the Parents and Children Together (PACT) programme on the language skills of 3- to 4-year-old nursery children, a randomised controlled trial [PACT-3]
DEVELOPER (INSTITUTION)	University of Manchester
EVALUATOR (INSTITUTION)	Durham University
PRINCIPAL INVESTIGATOR(S)	Victoria Menzies
SAP AUTHOR(S)	Bilal Ashraf, Jochen Einbeck, Victoria Menzies
Trial Statisticians	Professor Jochen Einbeck, Dr Bilal Ashraf
TRIAL DESIGN	Two-arm randomised controlled trial with random allocation at the child level within nurseries
Trial Type	Efficacy Trial
PUPIL AGE RANGE AND KEY STAGE	3–4-year-olds -EYFS
NUMBER OF SCHOOLS	43
NUMBER OF PUPILS	372
PRIMARY OUTCOME	Language skills assessed by LanguageScreen app at immediate post-test
SECONDARY OUTCOME	Language skills assessed by LanguageScreen app at 11 month delayed post-test Language skills assessed by researcher-delivered language assessments (BPVS, APT and CELF – Expressive Vocabulary) at immediate post-test and 11 month delayed post-test Specific domains of language skills (assessed by LanguageScreen subscale scores) at post-test and 11 month delayed post-test Early literacy skills as measured at 11 month delayed post-test using Early Word Reading, Letter-Sound Knowledge and Sound deletion subscale from YARC Child's home learning environment measured by the parent-reported Home Learning Environment Index at immediate post-test School readiness, measured by teacher- reported BESSI at immediate post-test

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]	n/a	

Table of contents

Study rationale and background

Vocabulary acquisition is a key element of early infant development and continues to be an important factor throughout childhood. Bergelson and Swingley (2012) reported that babies appear to start learning the sound forms of whole words within the first few months of life and they understand the meanings of several common nouns from the age of six months. At around the age of 18 months, young children's vocabulary begins to expand rapidly, and it is estimated that they learn words at a rate of one every two waking hours; a trend that will continue to adolescence (Pinker, 1994). In addition to vocabulary acquisition, infants need to learn about the features of spoken language such as where words begin and end and realise that these units carry a meaning. This phonological knowledge underpins vocabulary acquisition and growth.

Moving on to learning to read, Harrison (2004) suggested that children need different types of knowledge as precursors:

- Knowledge and understanding of the world; knowledge of how our language works.
- Knowledge of conventions of print; phonological awareness; decoding, oral reading fluency and reading comprehension are beginning to be acquired by many children by 5 years of age.

Evidence indicates that parenting and educational environment in the early years have a powerful influence on language development. The quality of the home learning environment and educational resources within the home are important factors (Melhuish *et al*. 2008b) and there is a link between this quality and socio-economic status (Foster *et al*. 2005). We observe children from disadvantaged backgrounds entering school with lower levels of attainment than their more socioeconomically advantaged peers (Tymms *et al*. 2014) and this trend persists throughout primary school (Merrell, Little and Coe, 2014); development and skills at the start of school are predictive of later outcomes (see, for example, Tymms, Merrell and Bailey, 2017).

Parents and Children Together (PACT) provides teaching sessions and materials for parents/carers to use with their children to develop their language skills. Previous research suggests that through its structured programme to teach children language skills and to enhance the home learning environment, PACT could positively impact on the quality of the home learning environment, leading to gains in language development and a previous trial of PACT delivered by the PACT developers (PACT-1; Burgoyne et al 2018) found that PACT had a positive impact on language skills which was still evident 10 months later. This trial is the third evaluation of the PACT programme (PACT-3) using a randomised controlled trial design. The second PACT trial (PACT-2) had its delivery significantly disrupted by Covid-19 restrictions and PACT-3 was commissioned before the PACT-2 trial had been completed.

While various dimensions of language skills are generally measured using standard tests, research has also shown that language skills reflect a unitary construct and can be measured using a latent variable sharing common variances of these dimensions (NELI Evaluation - Dimova et al., 2020; MacDonald, 2013). Latent variables have been used in assessing effectiveness of language interventions in different contexts, including assessment of parental interventions (Burgoyne et al., 2018). Selection of test components to construct a latent language skill variable is generally theory driven, but studies have used different combinations of these tests (Dimova, et al., 2020; Burgoyne et al., 2018; West et al.

2021). Recent research assessed correlation between two unitary constructs of latent language variables and observed high correlations between them (West et al., 2021).

INTERVENTION

PACT is an early language teaching programme delivered by parents or carers to their pre-school (aged 3-4 years) child in the year before they start school. It is an intensive programme delivered over a period of thirty weeks with focused language activities based on story books provided by the programme to be completed five days a week, for approximately 20 minutes a day. There are two levels of delivery in the programme with parents/carers signing up to and accessing the programme and programme support through their child's nursery and then delivering the programme sessions directly to their child at home. Training is provided at both levels for the nursery staff driving the programme (PACT Lead) and for the parents. In this trial the PACT programme is delivered by families between November 2021 and June 2022. PACT Lead training was provided in May 2021 and Parent/Carer training in November 2021.

Research questions

The research questions are focused on two time points:

- "Immediate post-testing" which takes place immediately after the intervention period when the children are aged 3-4 and still attending nursery.
- "Delayed post-testing" which takes place 11 months after the end of the intervention period when children are aged 4-5 and towards the end of their first year of formal schooling.

RQ1. What is the impact of the PACT intervention on language skills immediately after the intervention period, as measured by the LanguageScreen assessment? [Primary Outcome]

RQ2. What is the impact of the PACT intervention on language skills 11 months after the intervention period, as measured by the LanguageScreen assessment? [Secondary Outcome]

RQ3. What is the impact of the PACT intervention on the specific language domains of receptive vocabulary measured by the British Picture Vocabulary Scale (BPVS), expressive vocabulary measured by CELF Preschool 2 Expressive Vocabulary subscale (CELF-EV) and spoken language information and grammar measured by the Renfrew Action Picture test (APT Information, APT Grammar) immediately after the intervention period, using researcher-delivered assessments [Secondary Outcome]?

RQ4. What is the impact of the PACT intervention on the specific language domains of receptive vocabulary (measured by the BPVS), expressive vocabulary (measured by CELF-EV) and information and grammar in spoken language (measured by APT Information and APT Grammar) 11 months after the intervention period, using researcher-delivered assessments [Secondary Outcome]?

RQ5. What is the impact of PACT on school readiness immediately after the intervention period measured using teacher-completed Brief Early Skills and Support Index (BESSI)? [Secondary Outcome]

RQ6. What is the impact of PACT on the home learning environment as measured using the parent/carer-completed Home Learning Environment Index (HLE) at the end of the intervention period? [Secondary Outcome]

RQ7. What is the impact of PACT on early literacy skills as measured 11 months after the intervention period using the York Assessment of Reading Comprehension (YARC) assessment? [Secondary Outcome]

Research questions 1, 3, 5 and 6 will be investigated immediately following the intervention period at the end of nursery (immediate post-testing). Eleven months after immediate post-testing and the end of the intervention period, research questions 2,4 and 7 will be investigated while the participants are towards the end of their first year of formal school (delayed post-testing).

Study Design

The study design is a two-armed randomised controlled efficacy trial delivered under ideal conditions with allocation at pupil level. Pupils will be allocated equally to either the intervention (pupils allocated to receive the PACT programme) or control group (pupils allocated to 'business as usual' plus equivalent incentive cost of materials (approximately £130) in books to parents/carers on completion of the immediate post-test). The proposed within school randomisation will be more powerful than cluster randomisation if there is negligible heterogeneity in intervention effects between schools and if there is no dilution of the intervention effects as a result of contamination between intervention and control groups. The books included in the programme could be passed between parents/carers in the PACT group to the control group, however the developers advised the core of the programme is the accompanying activities and resources, which after initial completion would not be particularly useful as some of the materials are single use. Additionally, the staff training could encourage change of practice within the setting, however the developers intend the training to focus on the theory of the programme and how best to support parents/carers in its delivery, none of which is expected to create new knowledge significant enough to influence classroom-based practice. The summary of trial design is presented in Table 1. Unlike most EEF trials, the primary outcome is a latent construct of language development measured by a combination of LanguageScreen app sub-scale raw scores (Expressive vocabulary, Receptive vocabulary, Listening comprehension, and Sentence repetition). The latent variable will be obtained from a confirmatory factor analysis based on these sub-scores, similar to that done by West et al (2021). LanguageScreen is also being used in the national roll-out of the Nuffield Early Language Intervention (NELI) programme funded by the Department for Education (DfE) and the Education Endowment Foundation (EEF) to primary schools in England.

Table 1: **Study design summary**

Trial type and number of arms		**Two-armed randomised controlled efficacy trial**
Unit of randomisation		Pupil (within nurseries)
Stratification variables (if applicable)		Pre-test completeness, site
Primary outcome	variable	Language Skills (measured by app)
	measure (instrument, scale)	LanguageScreen, latent variable combining raw subscale scores (Expressive vocabulary 0-24, Receptive vocabulary 0-23, Listening comprehension 0-16, and sentence repetition 0-14), school delivered LanguageScreen app
Secondary outcomes	variable	Specific language domain skills of expressive vocabulary, receptive vocabulary and spoken language information and grammar (measured by researcher delivered assessments)
	measure(s) (instrument, scale)	1] Clinical Evaluation of Language Fundamentals Preschool 2 UK– Expressive vocabulary subscale (CELF-EV), 0-20, researcher delivered assessment [2] British Picture Vocabulary Scale – 3 (BPVS-3), 0-168 (raw score), researcher delivered assessment [3] Renfrew Action Picture Test (APT), information score 0 – 29, grammar score 0-38, researcher delivered assessment
	Variable	Specific language domain skills of expressive vocabulary, receptive vocabulary, listening comprehension and sentence repetition as measured by LanguageScreen subscales
	Measures (s)	LanguageScreen, raw subscale scores for [1]Expressive vocabulary (0-24), [2] Receptive vocabulary (0-23), [3] Listening comprehension (0-16) and Sentence repetition (0-14)
	variable	Early literacy skills
	measure (instrument, scale, source)	York Assessment of Reading for Comprehension (YARC) : [A] letter-sound knowledge core test raw score 0-17 researcher delivered assessment [B] early word reading test, raw score 0-30, researcher delivered assessment [C] sound deletion test, raw score 0-12, researcher delivered assessment
	variable	Home Learning Environment

	measure (instrument, scale, source)	Home Learning Environment Index (HLE), 0-49, parent/carer completed survey
	Variable	School Readiness
	measure (instrument, scale, source)	Brief Early Skills & Support Index (BESSI), 0-30 (total score), survey completed by nursery key worker
Baseline for primary outcome	variable	Language Skills (measured by app)
	measure (instrument, scale, source)	LanguageScreen, latent variable combining standardised subscale scores (Expressive vocabulary 0-24, Receptive vocabulary 0-23, Listening comprehension 0-16, and sentence repetition 0-14), school delivered LanguageScreen app at baseline
Baseline for secondary outcomes	variable measure (instrument, scale, source)	See Table 4 below for details of the baseline used for each secondary outcome

As in PACT-2, pre-test data will be collected at the start of the project. This will include data from the LanguageScreen measure and the Home Learning Environment outcome measure. The collection of pre-test assessment data from pupils will reduce the minimal detectable effect size and increase the power of the trial (see Table 3 below). The use of the LanguageScreen assessment both before the randomisation and again at the two post-testing time points will provide the strongest pre- post- test correlation (and greatest power to detect an effect). Home Learning Environment data will also be collected pre-test and immediate post-test.

Randomisation

The trial statistician who is not involved in the recruitment of nurseries or parents/carers, completed randomisation independently. All participating pupils were allocated into one of the two groups (intervention or control) on a 1:1 ratio. We applied a randomised block design, with nursery and pre-test subgroups serving as blocks, however with an additional element of paired randomisation to account for the presence of blocks of odd size. This is explained in full detail below.

Throughout, we will refer to a single sample in the study as an 'individual'. The study comprises 372 individuals across 43 nurseries (average cluster size=8.7).

The pre-test variable was originally coded as follows: 0 = not tested; 1 = tested; 2 = tried to test but child was shy/uncooperative. According to this categorization, 15 participants have pre-test status 0, 355 have pre-test status 1, and 2 have pre-test status 2. Since only two pupils had pre-test status 2, we

merged group 2 (members of which had an intention to be tested but the child did not engage with the assessment) with group 1 (actually tested) and consider both groups to have a pre-test status of 1 for further analysis. We were concerned that those that had not been tested at the time of randomisation may be systematically different from those where assessment had been conducted or attempted.

Our considerations in deciding a randomisation strategy were as follows, with decreasing priority:

- Every individual must have an equal probability of being assigned to the control or intervention group
- For any individual added to the study after the initial randomisation, study personnel should not be able to determine the group to which that individual will be allocated prior to the individual's enrolment. This is to avoid any confounding of results; if a school knows that their next recruit will be allocated to the intervention group, this may influence their decision to recruit another individual.
- We wish to allocate individuals to intervention and control groups in as close as possible to a 1:1 ratio
- We aim to allocate individuals to intervention and control groups to as close as possible to a 1:1 ratio within each nursery
- We aim to allocate individuals to intervention and control groups to as close as possible to a 1:1 ratio across pre-test statuses
- We aim to allocate individuals to intervention and control groups to as close as possible to a 1:1 ratio amongst individuals of the same pre-test status within each nursery
- We aim to allocate individuals to intervention and control groups to as close as possible to a 1:1 ratio amongst individuals at nurseries of similar size.

Given the considerations above, our choice of randomisation strategy is as follows:

1. We defined nursery-pre-test groups containing individuals in the same nursery and of the same pre-test status. For instance, nursery-pre-test group S02_1 contained individuals at nursery S02 with pre-test status 1.
2. For any nursery-pre-test group from which an even number of individuals are currently included in the study, we drew a random sample of half the size of the group without replacement from each such group and the individuals were split between control and intervention groups in equal number.
3. We then considered all nursery-pre-test groups containing an odd number of individuals. For each such group, we matched it with a second nursery-pre-test group of similar size and pre-test status. We did this by listing nursery-pre-test groups with pre-test status 0 by size and pairing the first off with the second, the third with the fourth, and so on, until there was one left. We paired this nursery-pre-test status group with the nursery-pre-test group with pre-test status 1 of the closest size. We then listed remaining nursery-pre-test groups with pre-test status 1 in ascending order of size, and matched the first with the second, the third with the fourth, and so on.
4. For each pair of nursery-status groups, we randomly assigned one of the pair to have one excess individual assigned to the control group and the other to have one excess individual assigned to the intervention group. We then stepped through pairings in ascending order of size to check if any nurseries had two odd-numbered groups both assigned to excess control or excess intervention (that is, one group with pre-test status 0 and another with pre-test status 1). When this occurred, we switched the pair assignments of the larger group and its pair.
5. For each nursery with an odd number of individuals favouring controls, assign one more individual at that nursery to the control group than to the intervention group, making assignments randomly across individuals. Make the corresponding assignment for schools favouring interventions.
6. For any additional individual added to the study, assign them randomly to either the control or intervention group with equal probability.

This strategy guarantees the following properties

- The 372 individuals in the study are split into control and intervention groups of equal size. (If ‘372 had to be replaced by an odd number, the size of control and intervention groups would differ by 1, with an equal probability that either group is larger.)
- In each nursery, the number of individuals in control and intervention groups differs by at most 1
- Whenever a nursery has 1 more control-group individual than intervention-group, a matched school with a similar number of individuals has 1 more intervention-group individual than control-group
- The 15 individuals with pre-test status 0 will be split into control and intervention groups of size differing by 1, with equal probability of the larger group being control or intervention
- The 355 individuals with pre-test status 1 will be split into control and intervention groups of size differing by 1, with equal probability of the larger group being control or intervention
- Within each nursery, the individuals with pre-test status 0 will be split into control and intervention groups of either equal size or size differing by 1, with equal probability of the larger group being control or intervention
- Within each nursery, the individuals with pre-test status 1 will be split into control and intervention groups of either equal size or size differing by 1, with equal probability of the larger group being control or intervention
- Individuals additional to the original 372 have an equal probability of assignment to each group, and the control/intervention assignment cannot influence the choice to recruit additional individuals.

Sample size calculations

The sample size calculations were done using Optimal Design software and they reflect a within school multi-site randomization design (Table 2). Our calculations assumed a 5% Type 1 error, 80% Power, 10% intra-school correlation, 60% pre-post-test correlation and used a two-sided test. The intra-cluster correlation of 10% is based on the average value observed in EEF trials (Xiao et al, 2016). The pre-post-test correlation value is that found in Burgoyne et al (2018) in a previous trial of the PACT intervention which used very similar language outcome measures to our primary outcome over the same period of time. Based on these assumptions, this sample size would detect a minimum difference of 0.18 standard deviations between the PACT and the control group (scenario 1a in Table 3 below).

Table 2: Sample size calculations

		OVERALL		EYPP
		At Design	At randomisation	
Minimum Detectable Effect Size (MDES)		0.18	0.21-0.22	0.72
Pre-test/ post-test correlations	level 1 (pupil)	0.60	0.60	0.60
	level 2 (class)	-	-	-
	level 3 (school)	-	-	-
Intracluster correlations (ICCs)	level 2 (class)	-	-	-
	level 3 (school)	0.1	0.1	0.1

		OVERALL		EYPP
		At Design	At randomisation	
Alpha[1]		0.05	0.05	0.05
Power		0.8	0.8	0.8
One-sided or two-sided?		Two-sided	Two-sided	
Average cluster size		10	8 – 9	
Number of schools[2]	Intervention	48	43	-
	Control	48	43	-

Table 3 below also explores varying the pre-post-test correlation to look at the potential impact of not including a baseline assessment (scenario c with Pre-post correlation of 0) or including a less correlated measure in the analysis (scenario b with Pre-post correlation of 0.3). Investigation of these figures led us to the decision that it was necessary to include a well correlated covariate in the analysis and the using the same assessment at pre-test gave the trial the power necessary to detect the level of effect size found in the previous Burgoyne et al 2018 trial.

Table 3. MDES using a variety of pre-post-test correlation assumptions and varying recruitment levels of schools and pupils

Scenario	Significance level (α)	Power (1— β)	Effect size variability estimate	Pre-post correlation (R^2)	ICC	no pupils per school (n)	MDES if 43 schools (J)	MDES if 44 schools (J)	MDES if 45 schools (J)	MDES if 48 schools (J)	MDES if 50 schools (J)
1a	0.05	0.8	0.05	0.6	0.1	10	0.20	0.19	0.19	**0.18**	0.18
1b	0.05	0.8	0.05	0.3	0.1	10	0.25	0.25	0.24	0.23	0.22
1c	0.05	0.8	0.05	0	0.1	10	0.29	0.29	0.28	0.27	0.26
2a	0.05	0.8	0.05	0.6	0.1	9	**0.21**	0.20	0.2	0.19	0.19
2b	0.05	0.8	0.05	0.3	0.1	9	0.26	0.26	0.25	0.24	0.23
2c	0.05	0.8	0.05	0	0.1	9	0.31	0.30	0.29	0.28	0.27
3a	0.05	0.8	0.05	0.6	0.1	8	**0.22**	0.21	0.21	0.2	0.2
3b	0.05	0.8	0.05	0.3	0.1	8	0.28	0.27	0.26	0.25	0.25
3c	0.05	0.8	0.05	0	0.1	8	0.33	0.32	0.31	0.29	0.29
4a	0.05	0.8	0.05	0.6	0.1	7	0.23	0.22	0.22	0.21	0.21
4b	0.05	0.8	0.05	0.3	0.1	7	0.29	0.28	0.26	0.25	0.25
4c	0.05	0.8	0.05	0	0.1	7	0.34	0.34	0.32	0.31	0.31

EYPP

This trial is not powered to detect an effect size on the sample of children who are eligible for Early Years Pupil Premium (EYPP) and schools have not been able to share information on the EYPP of children in the sample so far. It is planned to collect this data in June 2022 at the end of the nursery year. The current figures in above Table 2 are expected numbers calculated (in optimal design

[1] Please adjust as necessary for trials with multiple primary outcomes, 3-arm trials, etc., when a Bonferroni correction is used to account for family-wise errors.
[2] Please adjust as necessary, e.g., for trials that are randomised at the class level.

software) based on the assumption of 16% EYPP pupils as we had in PACT 2. This table will be updated once data will be available.

Outcome measures

Baseline measures

LanguageScreen (LS) Assessment

To provide adequate power to the research it is necessary to use a baseline measure which correlates well with the primary outcome. This study is using the same language skills assessment at baseline as for the primary outcome. This gives the best chance for a good correlation (assumed 0.6 as described above in sample size section) between baseline and immediate post-test. The LanguageScreen assessment is therefore used with all participants in September 2021 before randomisation and the LS latent variable score at baseline will be used in the primary analysis. Further details of LanguageScreen are included below in the Primary outcome section.

Where the LS subscales closely align with the language secondary outcomes measures, the aligned subscale completed at baseline will be used in the secondary analysis. Specifically, the LS Receptive Vocabulary (LS-RV) subscale raw score will be used as baseline for the BPVS outcome, and the LS Expressive Vocabulary (LS-EV) subscale will be used as baseline for the CELF Expressive Vocabulary (CELF-EV). Where no specifically aligned subscale is available for the secondary language outcomes, the baseline LS latent variable score will be used as the baseline score. Full details are included in Table 4 below.

Home Learning Environment

The Home Learning Environment Index (HLE; Melhuish et al, 2008a) was developed as part of the EPPE study and has been used in several large studies including the Millennium Cohort Study, National Evaluation of Sure Start (NESS) and a study of the Home Learning Environment by the Scottish Government (Melhuish, 2010). The HLE asks parents/carers to report the frequency of seven routine activities which are conceptually linked to learning (including being read to, going to the library, playing with numbers, painting and drawing, being taught letters, being taught numbers and songs/poems/rhymes). These seven items were positively linked with predicting under and over achievement at aged 5 (Melhuish et al. 2008a, 2008b). Frequency of the seven activities is coded on a 0 to 7 scale with 0 representing “not at all” and 7 representing high frequency for the activity (varying depending on the activity), yielding a total score of between 0 and 49 with higher scores indicating a a higher quality home learning environment. The Home Learning Environment is a secondary outcome, and the baseline measure total raw score will be used as a covariate in the secondary outcome analysis.

PRIMARY

The primary outcome will be language skills as measured using the LanguageScreen assessment (delivered by OxEd Assessment: https://oxedandassessment.com/language_screen). Language skills

including improved vocabulary is one of the key outcomes that PACT aims to improve, and improvements have been shown to a language skills latent variable outcome in the previous trial of PACT (Burgoyne et al 2018). Specific use of LanguageScreen as the primary outcome in this trial is for a number of reasons. Firstly, LanguageScreen measures four aspects of language skills giving a broad measure of language skills and two subscales are specifically aligned to the intervention's particular focus on vocabulary. Secondly, the measure is delivered by school staff in the classroom and doesn't require external researchers to be able to visit the school. With varying and ongoing covid-19 restrictions this seems to give a good likelihood of being able to collect data even if researchers would be unable to visit the schools. Thirdly, as it is delivered by school staff and schools are incentivised to assess children then it is hoped that there may be less attrition due to children being absent on the day of assessment. LanguageScreen has been used to collect assessment data in PACT-2 and for schools that were involved in the project at the time of post-testing there was a good LanguageScreen response rate (96%).

LanguageScreen is a standardised app-based assessment which is delivered by a member of school staff. The LanguageScreen assessment (delivered by OxEd assessment: https://oxedandassessment.com/language_screen) is made up of four subscales: Receptive Vocabulary (23 items where the child chooses which of 4 pictures matches a spoken word (raw score range 0-23)), Expressive Vocabulary (24 items asking the child to name pictures (raw score range 0-24)), Listening Comprehension (listening to 3 stories each followed by a series of questions about the story tapping understanding – 16 items (raw score range 0-16) and Sentence Repetition (14 items repeating verbatim a series of spoken sentences (raw score range 0-14)). It is administered using an app on a tablet by a member of staff in the child's school. Full instructions are included within the app for the delivery of the assessment without the need for training. Verbal instructions and items for the child are played aloud through the app. This should minimise variability in the delivery of the assessments across all the settings. The four assessments are presented in a set order and take around 25 minutes to complete. The assessment administrator marks on the app whether the child answers correctly or not for questions where the child gives a verbal answer. Data from the app is uploaded to the LanguageScreen website automatically and results are generated automatically by LanguageScreen. A standardised and raw score for each subscale as well as overall raw and standardised scores is provided. We will use a latent variable (West et al. 2021) formed from the four subscales raw scores. The latent variable will be a weighted sum of these raw scores, with weights obtained from a confirmatory factor analysis.

Model fit will be assessed using the following criteria: root mean square error of approximation (RMSEA) <0.08; standardized root mean square residual (sRMR) <0.08; comparative fit index (CFI) ≥ 0.90; and Tucker-Lewis index (TLI) ≥ 0.95 (Hu and Bentler, 1999; Kline, 1998). The weightings obtained from the post-test CFA will be applied to the baseline scores too, even though a CFA will also be carried out on the baseline scores to check for the consistency of the weightings obtained. If the planned model is not a good fit, or if major inconsistencies arise in this process, modifications necessary will be made in consultation with the EEF.

The primary outcome measure will therefore be a latent variable created using LanguageScreen raw subscale scores at immediate post-test. A similar modelling approach was used in previous PACT trials (Burgoyne et al., 2018).

LanguageScreen assessment scores correlated strongly (r=.95) with a latent variable created from scores on standardised researcher-delivered measures (CELF-EV subscale and APT information and grammar scores as well as CELF-Preschool Recalling Sentences) in a previous study of more than a thousand participants (West et al 2021). This gives a strong indication that the LanguageScreen assessment is measuring the same constructs as the latent variable in the West et al study and in the previous PACT trial (Burgoyne et al 2018) and should be a good measure for this research given the similarity of assessments.

Schools will deliver LanguageScreen using the LanguageScreen app on a tablet or large phone. The app guides the assessment, reading aloud all the questions and text. The adult in the school supporting the assessment delivery decides whether the child has responded correctly or incorrectly and presses the appropriate button. There is guidance in each section for the adult delivering the assessment. Assessors using the app will be encouraged to use a practice code to do a run through of the assessment in advance of assessing any children. The assessment takes between 10 and 20 minutes to complete for each child. It will not be possible to blind the assessor to the intervention allocation of the child, therefore there is the potential for bias in the completion of the assessments. The app delivery of the assessment is very structured and leaves little room for varying the delivery of the assessment, so we expect potential ascertainment bias to be minimised. It was decided to use LanguageScreen as the primary outcome to mitigate for the risk that Covid-19 restrictions may mean it is impossible for researchers to visit schools to collect data, and the small potential risk of bias was deemed acceptable. It also allowed for the primary outcome to be collected independently of the developer team with data processing and scoring being automated by the LanguageScreen assessment process again minimising the risk of ascertainment bias.

For the delivery of LanguageScreen, the developer team will prepare the data for the participating pupils and do initial communication with schools about the upcoming assessment period. The evaluator team will upload all the pupil information into the assessment software and will liaise with schools during the testing period to support their delivery of LanguageScreen. Should schools have difficulties with accessing the LanguageScreen assessment on their hardware the evaluation team will be able to courier tablets to schools for schools to conduct the assessments.

SECONDARY

Language skills at delayed post-test

Participants will be assessed again at delayed post-testing 11 months after the intervention delivery period using the LanguageScreen assessment again. The LanguageScreen data will be collected in the same way as for the primary outcome and the same latent variable will be used in the analysis. There is less potential for bias as the staff delivering the assessment when the child is in reception is less likely to know whether the child received the PACT intervention or not. In the previous Burgoyne

et al (2018) trial of PACT, larger effect sizes were found on language skills at delayed post-test, and we are testing the hypothesis that the same will be found in this trial. This would indicate that the PACT intervention has lasting effects on children's language skills if the same finding were obtained in this trial.

Specific domains of language skills

LanguageScreen subscales scores of Receptive Vocabulary (LS-RV), Expressive Vocabulary (LS-EV), Listening Comprehension (LS-LC) and Sentence Repetition (LS-SR) as collected at immediate and delayed post-test will be used as individual secondary outcomes. The logic model expects that PACT will impact on the whole language skills of the child however different programme activities are targeted towards specific language domains including listening comprehension, vocabulary, narrative skills and sentence level language skills. The raw subscale scores of each domain of Language Skill will therefore be used to investigate whether there is a greater improvement in these specific domains of language skill.

Researcher delivered Language Skill Measures

The raw scores on three measures of language skills in the domains of expressive vocabulary, receptive vocabulary and spoken language information and grammar, delivered by researchers face-to-face in schools, will be used to investigate impact in these specific domains of language skill as well as to triangulate with the primary outcome measure and to investigate whether the school delivery of LanguageScreen introduces bias due to knowing the assignment of the child being assessed:

(a) The British Picture Vocabulary Scale – 3 (BPVS-3), is a standardised measure of receptive vocabulary appropriate to 3-year-olds up to adult. The programme activities specifically target vocabulary learning and involve increased exposure to a variety of books and resources. This measure consists of a set of pictures from which the child is asked to point to the picture representing a given word. This assessment gives a raw score between 0 and 168 with a higher score indicating a wider receptive vocabulary and lower score indicating a narrower receptive vocabulary.
(b) The Clinical Evaluation of Language Fundamentals Preschool 2 UK expressive vocabulary subscale score (CELF-EV). This specific subscale provides a measure for expressive vocabulary skills in young children. This is a standardised and validated assessment with the proposed age group and UK sample and has been used in previous studies to look at the impact of PACT (Burgoyne et al., 2018) as well as other EEF funded Early Years Studies (e.g.,https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/easypeasy-learning-through-play). This assessment gives a raw score between 0 and 20 with a higher score indicating a child has a wider expressive vocabulary and a low score indicating a narrower expressive vocabulary.
(c) The Renfrew Action Picture Test (APT) is a standardised test that requires children to give samples of spoken language in response to picture stimuli. The test considers grammatical structures used and the expressive vocabulary used. The test is suitable to use with children between the ages of 3 and 8 and provides normed scores. It is quick and simple to administer and inexpensive to purchase. This assessment provides two scores – information score (range

0-29) and grammar score (0-38). A higher score on these scales indicates more advanced spoken language skills in terms of the vocabulary used for the information score and the grammatical structures used for the grammar scores. The grammar aspect of this assessment is not captured by the primary outcome LanguageScreen and is another aspect of language skills we feel is important to capture for this trial. Both scores will be considered separately.

These assessments will be conducted at immediate post-test right after the intervention delivery period (end of nursery) and at delayed post-test 11 months later.

The delivery of these researcher delivered measures will be the responsibility of the developer team who will recruit and provide training to a team of Research Assistants (RAs) to collect this assessment data. Training for RAs will consist of an off-site training session day plus additional on-site training where the RA observes the delivery of two assessments by a member of the core-developer team (not blinding to allocation) and then the developer observes the delivery of the assessment by the RA. The evaluator team will quality assure the delivery of an audio-recorded practice assessment by each RA before any onsite visits. This will involve assessing the delivery of each assessment according to a pre-agreed protocol and providing feedback to the developer team for each RA on the quality of the assessment delivery. Both the evaluator quality assurance and the onsite observation of each RAs first assessment will quality control delivery and allow immediate feedback to be given to RAs if required.

The RAs will be blinded to the allocation of the child. The assessments will be conducted in the pre-specified order of BPVS, CELF-EV and the Action Picture Test and will take approximately 15-20 minutes to complete per child. Scoring and data entry of these assessments will also be done by the core research team; the use of identifier codes, and allocation data being kept separate to the outcome measures will minimise the potential balance from not being completely blinded. Parents/carers will be asked to give permission for the RA delivered assessments to be audio recorded. All RA delivered assessments, where permission has been given, will be audio recorded for the evaluation team to perform quality assurance on 10% of the assessments carried out by the RAs, and all of the assessments carried out by the developer team. We will randomly select 10% of the assessment conducted by each assessor for the quality assurance process excluding their first two assessments which were observed in person by a member of the developer core team. Raw data scores will be securely transferred to the evaluation team for independent analysis.

Sensitivity analyses will be conducted to investigate differences on language scores through LanguageScreen and those collected by blinded researchers visiting the schools.

Early literacy skills

Early Literacy Skills (measured at the 11-month post-intervention delayed post-test) will involve measures of Letter-Sound Knowledge, Early Word Recognition and Sound Deletion (from the YARC Early Reading assessment). For the Letter-Sound Knowledge test the child is presented with lower case letters and digraphs, one at a time, and is required to say what sound the letters and digraphs make. The core test of letter-sound knowledge will be used giving a raw score of between 0 and 17 with a higher score showing greater letter-sound knowledge. For the Early Word Recognition test the child

is shown up to 30 words graded in difficulty and asked to say what the word is, giving a raw score of between 0 and 30 – a higher score indicates the child can read a greater number of words. In this assessment, half of the words are phonemically regular (can be decoded) and half of the words are phonemically irregular (exception) words and raw subscale scores will be generated for regular and irregular early word reading (0-15 for each subscale) and also used as secondary outcomes. For the Sound Deletion test, the child hears a word (and sees a picture of the word) and is asked to repeat the word with a sound taken away. The test consists of 12 items and the child is given a raw score from 0 to 12 – a higher score shows greater phonetic awareness. The scores of each scale will be presented separately.

In the previous trial of PACT, the largest effect sizes were found at delayed post-test on these measures which are indicators of early literacy skills rather than the pre-literacy language skills measured by the other outcomes. For the Early Word Reading significant effects were found specifically for the phonemically regular words and not the irregular words and we will investigate if this is replicated here. The York Assessment of Reading Comprehension (YARC) Early Reading assessment is a standardised and validated measure of alphabetic knowledge, single word reading and phonemic awareness skills and is particularly appropriate for use in the 4-to 6-year-old age group. This assessment will only be used at delayed post-test as the assessment material is not relevant before this stage. This measure aligns with the logic model with the expectation that improved language skills and school readiness will lead to improved early literacy skills. These assessments will be conducted by the same research assistants at the same time as administering the researcher-delivered language skills assessments at delayed post-test. The RAs will be blinded to the trial allocation of the child. This data will be collected and scored by the developer team and transferred to the evaluation team via secure file transfer. The developer team will be responsible for the marking of the data, but this will be conducted blind to the participant and trial allocation.

School Readiness

The Brief Early Skills & Support Index (or BESSI) questionnaire (measured at the immediate post-test) (Hughes et al, 2015) will be used to evaluate school readiness which is expected to improve as a result of the PACT programme. BESSI is a standardised 30 item questionnaire which assesses how well children are making the transition to school. This questionnaire has been developed and validated for reception and nursery children. Questions are to be answered for an individual child to reflect the child's behaviour over the previous week and statements are answered on a four point strongly agree to strongly disagree scale. Some items are reverse worded, and the four-point scale is converted to 2 points (agree or disagree) for scoring. This scale contains 4 subscales measuring Behavioural Adjustment (BESSI-BA; 12 items), Language and Cognition (BESSI-LC; 6 items), Daily Living Skills (BESSI-DLS; 6 items), and Family Support (BESSI-FS; 6 items). A lower score indicates a greater level of school readiness in each scale while a higher score indicates the prevalence of more problematic behaviours for school. This study will use each of these subscales separately to look at the different aspects of school readiness as well as total score so that the results could be compared to the PACT-2 study.

For the delivery of BESSI, school PACT Leads will be emailed a link to an online survey which should be completed for all the participating students in their settings. They will be asked that a member of staff who knows the child completes the BESSI for that child (ideally the child's keyworker). For settings that have difficulties accessing the online survey, a paper copy will be provided that they can copy for each pupil and return by post. Instructions for completing BESSI will be sent at the same time as those for LanguageScreen at immediate post-test.

Home Learning Environment

The Home Learning Environment Index (HLE; Melhuish et al, 2008a) described above in the Baseline Measures section will be collected at immediate post-test as part of the post-intervention usual practice surveys. Surveys will be distributed to parents/carers via email addresses in the first instance or to home addresses if the parent has not provided an email address. A prize draw incentive of four Amazon vouchers will be included to encourage participation in these surveys.

Statistical Analysis

The primary outcome and secondary outcomes will be analysed using the principles of intention to treat, meaning that all schools and pupils will be analysed in the group they were randomised to, irrespective of whether or not they actually participate in the PACT programme. Statistical significance will be assessed at the 5% level. Point estimates and 95% confidence intervals will be provided as appropriate. Results will be reported according to the EEF reporting template.

Imbalance at baseline

All the baseline data will be presented by intervention and control group using descriptive statistics. Cross-tabulation of background characteristics (including gender and by Early Years Pupil Premium (EYPP)) will be presented. We will also perform cross-tabulation between the pre-test status (Pre-test completed, pre-test not completed at time of randomisation and No pre-test data available due to being uncooperative) and the intervention status. Additional data on pupils and school characteristics will be described. For continuous variables, we report means and standard deviations of raw scores and standardised scores where available, and for categorical data, counts and percentages. Note that no effect size will be presented for baseline data.

Primary Outcome

The primary outcome is a latent language variable derived by combining four variables from scores on LanguageScreen subscales (1. Expressive vocabulary, 2. Receptive vocabulary, 3. Listening comprehension, 4. Sentence repetition). The weightings of this latent variable will be extracted from the loadings matrix of a confirmatory factor analysis based on the four raw Language Screen subscales, and will be considered as fixed and known henceforth. The latent variable is then obtained as weighted sum according to these weights.

The outcome variable will be analysed using a multilevel model. The pre-test LanguageScreen language latent variable (formed from the assessment collected at pre-test) will be included as a covariate for baseline adjustment. The effect size and its confidence/credible intervals will be computed using unconditional variance of the outcome data by fitting the multilevel models using the R package eefAnalytics (Robust Analytical Methods for Evaluating Educational Interventions using Randomised Controlled Trials Designs). School and school-by-intervention will enter as random effects into the multilevel model.

The analyses of outcomes follow an Intention to Treat (ITT) principle, as suggested by the EEF Statistical Analysis guidelines. Since the study was a multisite trial, we use multilevel models (MLM) adjusted for prior attainment which will account for the variability in average pupil attainment across schools participating in the trial and variation in the intervention effect across schools. The choice of analytical model is considered an optimal choice following the study design. The model specification for the empty unconditional model (required for effect size denominators) and for the conditional model including intervention and pre-test as a covariate is shown below.

$$y_{ij} = \begin{cases} \beta_{00} + b_{0j} + \epsilon_{ij0} & \text{for unconditional model} \\ \beta_0 + \beta_1 t_{ij} + \beta_2 pretest_{ij} + b_{1j} + b_{2j} t_{ij} + \epsilon_{ij} & \text{for conditional model} \end{cases}$$

Here, y_{ij} = Outcome variable (continuous), for ith child in jth school where $j = 1,2,\ldots,M$ and $i = 1,2,\ldots,n_j$.
M = number of schools,
n_j = number of children in each school,
$\epsilon_{ij} \sim N(0, \sigma_1^2)$ = conditional residual error,
($\epsilon_{ij0} \sim N(0, \sigma_0^2)$) = unconditional residual errors reflecting individual child differences in post-test, and
t_{ij} is = intervention variable for child i in school j.
$b_{1j} \sim N(0, \sigma_{11}^2)$, $b_{2j} \sim N(0, \sigma_{22}^2)$ = random effects capturing the variation between schools from conditional models, and $b_{0j} \sim N(0, \sigma_{00}^2)$ = random effects from unconditional models.
β_1 = regression coefficient for the intervention variable for child i in school j.
$pretest_{ij}$ = pre-test variable for child i in school j.
β_2 = the regression coefficient for the pre-test variable for child i in school j.

The previous PACT2 trial was investigated with a similar approach. This modelling approach enables estimation of impacts of PACT across the different domains of language skills as measured by the latent outcome. It assumes that the language skills may be better assessed as a latent construct that uses shared variance of the subscales and can reflect important elements of language skills that may be difficult to measure relying on observed variables. This multilevel approach also allows us to test whether the estimated effects of the intervention are constant across schools.

Sensitivity Analyses

The raw LanguageScreen subscale-scores used to produce the latent variable are not age-adjusted (unlike the standardised ones). As a consequence, the latent variable that will be produced from the confirmatory factor analysis will not be age-adjusted either. While it is acknowledged that the spanned age range is narrow (a year), at such a young age a difference of several months could still make a significant difference. Hence, it is a valid question whether the inclusion of age (months since the third birthday) into the multilevel model will make a notable difference in the analysis. Our expectation is that

this age variable will be highly correlated with the pre-test (baseline) score, and that therefore its impact will be limited. The sensitivity analysis will compute and report this correlation, and also re-fit the multilevel model including age and report the outcome.

Additional sensitivity analysis will be carried out to investigate whether there are differential effects of the intervention depending on the baseline score. This will be achieved by adding an interaction term intervention/pre-test to the multilevel model.

Secondary Outcomes

All non-latent variable secondary outcomes will be analysed using multilevel models with school and school-by-intervention as random effects. The effect size and the associated confidence/credible intervals will be calculated using unconditional variance of the outcome data to ensure consistency of results with the latent variable model, where the confidence/credible interval for the effect of the intervention will be based on unconditional variance. The immediate (HLE, BESSI-BA, BESSI-LC, BESSI-DLS, BESSI-FS, BPVS, CELF-EV, APT, LS-RV, LS-EV, LS-LC, LS-SR) and delayed impacts (LanguageScreen (as assessed using a latent variable approach), BPVS, CELF-EV, APT, YARC-LSK, YARC-PA, YARC-EWR (total, regular words and irregular words), LS-RV, LS-EV, LS-LC, LS-SR) of the PACT intervention on the secondary outcomes will be analysed using a multilevel model accounting for intra-school correlation (Further guidelines are available in the EEF 2018 Statistical Analysis guidance document). Where available, an appropriate pre-test variable will be included in the model as a covariate for baseline adjustment (see Table 4 below for specific details).

Table 4. Baseline covariates to be used for secondary outcomes

Secondary outcome	**Pre-test variable to be included**
HLE	HLE at pre-test
BESSI (all subscales)	-
BPVS	LS-RV at pre-test
CELF-EV	LS-EV at pre-test
APT	LS latent variable at pre-test
YARC – EWR (total, regular & irregular words)	LS latent variable at pre-test*
YARC - LSK	LS latent variable at pre-test*
YARC - PA	LS latent variable at pre-test*
LS-RV	LS-RV at pre-test
LS-EV	LS-EV at pre-test
LS-LC	LS-LC at pre-test
LS-SR	LS-SR at pre-test

*While the YARC is measuring aspects of literacy which are not directly comparable with the baseline language, it is expected that language is foundational for reading and that the addition of the baseline covariate will provide additional power for the analysis.

SUBGROUPS ANALYSIS

All the outcome data will be analysed by Early Years Pupil Premium (EYPP) eligibility model. Alongside fitting the latent variable model separately for subgroup category of those who received EYPP, interaction model will also be considered. Effect size for pupils eligible for EYPP will be reported in accordance with EEF requirement.

Longitudinal follow-up analyses[3]

No longitudinal data collection has been planned. The analysis dataset (excluding LanguageScreen results) will be archived, and longitudinal analysis could be conducted in the future by linking the data with NPD. LanguageScreen delayed post-test data will be pseudonymised when archived and will not be available to link with NPD as there is no agreement obtained with the LanguageScreen developer to link the LanguageScreen dataset with administrative data.

Missing data

Missing data may occur in the pre-test and outcome (post-test) measures and will be assessed at both. Pre-test missing data will be presented using cross-tabulation between missing data completeness status (completed, partially completed and no pre-test data) and the intervention groups. We would also investigate percentage of missing data in each of the individual components of the primary outcome data, and further analysis regarding imputations will be done when >5% of the outcome data are missing. The latent variable approach with the full information maximum likelihood estimation used for the CFA step implicitly assumes that the underlying mechanism for the missing data does not depend only on the observed data. This missingness mechanism is commonly termed 'missing at random' (MAR) and the full information maximum likelihood method (Cham et al., 2017) estimates the parameters of the latent variables conditioning on the observed data for each of the latent outcomes. It also assumed that all outcomes are linearly related with each other and are multivariate normally distributed, which enables it to condition missing data on observed data assuming multivariate normal distribution (Charm et al., 2017). In order to check whether the assumption of MAR holds, we would also perform multiple imputation on the composite outcomes and then apply latent variable model to estimate the impact of the intervention. We would expect that results from multiple imputation and the full information maximum likelihood estimation led to similar conclusion if the underlying missingness mechanism is missing at random. We would consider ten imputations for each outcome using chained equations or the Markov chain Monte Carlo (MCMC) method, which allows non-monotone imputation between pre-test and post-test data (Jakobsen et al., 2017). To impute pre-test data for a particular outcome, the pre-test scores for other outcomes will be used in the imputation model. However, both pre-test and post-test data will be used in the imputation model for any of the post-test outcomes. The imputation approach will be sequential such that all the pre-test scores will first be imputed and then they would be used in turn to impute the post-test outcomes. Note that the effect size from each of the imputation will be presented as range of values for sensitivity analysis. We would not consider a dropout model for the multiple imputation because of the nature of the latent variable model. The collective

missing data is more important in the latent variable model rather than individual dropout model. Lastly, we would use all the available data on the latent primary outcome for missing data imputation.

CACE Analysis

The self-reported compliance data will measure to what extent each child in the intervention group adhered to the required sessions of the intervention. Compliance data on number of sessions delivered based on data submitted by parents about each session using the PACT app and paper record forms (total number of sessions completed) will be used in a Complier Average Causal Effect (CACE) analysis. The CACE analysis will be implemented using an instrumental variable approach by comparing the outcomes between the intervention group and control group with a focus on random variation in compliance data. In other words, it will assess when conditioning on the number of PACT session/home support, what is the impact of the PACT intervention on language development (Pokropek, 2016). The model will assume that the intervention assignment impacts the outcome only via compliance. This means that there would be no impact of the intervention if a child had no session of PACT. Under this assumption, instrumental variable would be incorporated as an additional node in a structural equation model. However, the instrumental variable will be considered observed instead of latent with the disadvantage that it does not allow for testing whether the self-reported compliance data is a valid instrument (Pokropek, 2016). CACE analyses will also be carried out using a sequence of binary cut-offs, including 50 and 80% of completed sessions, to define an intervention as `compliant'

Intra-cluster correlations (ICCs)

There is no explicit estimation of ICCs in a latent variable model. However, we will estimate ICCs for the analysis of the individual outcome data using multilevel models. The pre-test estimation of ICCs will be based on a model with only the overall mean and with schools as random effects. The estimation of ICCs for post intervention data will be done with and without fixed effects, but with schools as random effects. ICCs will be computed at school level.

Effect size calculation

The effect sizes for the primary and secondary outcomes will be obtained from fitted multilevel models, using Hedges' g effect size defined as

$$ES = \frac{\widehat{\mu_T} - \widehat{\mu_C}}{\sqrt{\sigma_w^2 + \sigma_s^2 + \sigma_I^2}}$$

Where $\widehat{\mu_T} - \widehat{\mu_C}$ is the adjusted average difference between the intervention and control groups. σ_w^2 is residual variance, σ_s^2 denotes between school variance and σ_I^2 denotes the variance of school by intervention effects. As per analysis guidelines by the EEF, our main analysis for estimating effect size used the unconditional variance generated from an empty model in the denominator, while estimates for the numerator of the effect size is obtained from the conditional multilevel model. We will also compute the effect size using conditional variance.

Software

We will analyse the data using statistical software R (most updated version at the time of analysis).

Reporting of results

All results will be reported using the EEF template as shown in Table 4.

Table 5: Template for reporting results

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges g (95% CI)	p-value

Cost evaluation

Data on intervention costs will be collected from the developers as well as from schools participating in the PACT Lead interviews, through interviews as part of the process evaluation, and will be used to conduct a cost evaluation in line with recent guidance from the EEF.

Data protection

The legal basis for processing the personal data accessed and generated by the trial is Public Task covered by GDPR Article 6 (1) (e) public task, which states that; “the processing is necessary for you to perform a task in the public interest or for your official functions, and the task or function has a clear basis in law.” No special category data will be collected as part of this project.

References

Bergelson, E. & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns, *PNAS*, *109* (9), 3253 – 3258. www.pnas.org/cgi/doi/10.1073/pnas.1113380109 Article retrieved on 06/06/12

Burgoyne, K., Gardner, R., Whiteley, H., Snowling, M.J. & Hulme, C. (2018). Evaluation of parent-delivered early language enrichment programme: evidence from a randomised controlled trial. *The Journal of Child Psychology and Psychiatry*, *59*, 5, pp.545-555.

Cham, H., Reshetnyak, E., Rosenfeld, R. & Breitbart, W. (2017). Full Information Maximum Likelihood Estimation for Latent Variable Interactions with Incomplete Indicators. *Multivariate Behavioral Research*., *52*(1): 12–30. doi:10.1080/00273171.2016.1245600.

Dimova, S., Ilie, S., Rosa Brown, E., Broeks, M., Culora, A., & Sutherland, A. (2020). *The Nuffield Early Language Intervention Evaluation report.* Education Endowment Foundation. Available at https://educationendowmentfoundation.org.uk/public/files/Nuffield_Early_Language_Intervention.pdf

Jakobsen, J.C., Gluud. C., Wetterslev, J & Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. *BMC Medical Research Methodology, 17*:162. DOI 10.1186/s12874-017-0442-1

Harrison C. (2004) *Understanding Reading Development*, Sage Publications, London.

Hu, L. t. & 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.

Kline, R. B. 1998. *Principle and practice of structural equation modeling*, New York, Guilford.

MacDonald, M. (2013). How language production shapes language form and comprehension, *Frontiers in Psychology*, 4, 226 https://doi.org/10.3389/fpsyg.2013.00226

Melhuish, E.C., Sylva, K., Sammons, P., Siraj-Blatchford, I., Taggart, B., & Phan, M. (2008a). 'Effects of the Home Learning Environment and preschool center experience upon literacy and numeracy development in early primary school'. *Journal of Social Issues*, 64, 157-188.

Melhuish, E.C., Sylva, K., Sammons, P., Siraj-Blatchford, I., Taggart, B., Phan, M., & Malin, A. (2008b). 'Preschool influences on mathematics achievement'. *Science*, 321, 1161-1162 .

Melhuish, E.C. (2010). *Impact of the Home Learning Environment on Child Cognitive Development: Secondary Analysis of Data from 'Growing Up in Scotland'.* Report available on the Scottish Government Website. http://www.gov.scot/Publications/2010/04/27112324/0

Merrell, C., Little, J. & Coe, R. (2014) *Is the Attainment Gap among Primary Aged Children Decreasing? In Harnessing what works in eliminating educational disadvantage: A tale of two classrooms*, Eds. Wood, C. and Scott, R. Pub. Demos: London.

Pinker, S. (1994a). The language instinct. New York: Morrow.

Pokropek, A. (2016) Introduction to instrumental variables and their application to Large-scale assessment data. *Large-scale Assess Educ*, 4:4. DOI 10.1186/s40536-016-0018-2

Tymms, P., Merrell, C., Hawker, D., & Nicholson, F. (2014). *'Performance indicators in primary schools: A comparison of performance on entry to school and the progress made in the first year in England and four other jurisdictions'*. Research Report for the Department for Education.

Tymms, P., Merrell, C. & Bailey, K. (2017). The Long Term Impact of Effective Teaching. *School Effectiveness and School Improvement.* Published online, 14th December 2017 http://dx.doi.org/10.1080/09243453.2017.1404478)

West, G., Snowling, M. J., Lervåg, A., Buchanan-Worster, E., Duta, M., Hall, A., ... & Hulme, C. (2021). Early language screening and intervention can be delivered successfully at scale: evidence from a cluster randomized controlled trial. *Journal of Child Psychology and Psychiatry*.