Impact Evaluation of Nuffield Early Language Intervention Wave 2
Evaluation Study Plan

Evaluator (institution): The National Foundation for Educational Research
Principal investigator(s): Jack Worth

PROJECT TITLE	**Impact Evaluation of Nuffield Early Language Intervention (NELI) Wave 2**
EVALUATOR (INSTITUTION)	National Foundation for Educational Research
PRINCIPAL INVESTIGATOR(S)	Jack Worth
STUDY PLAN AUTHOR(S)	Jack Worth, Andrew Smith, Aarti Sahasranaman, Ruth Staunton
STUDY DESIGN	Regression Discontinuity
PUPIL AGE RANGE AND KEY STAGE	4-5 years; Reception
NUMBER OF SCHOOLS	548[1]
NUMBER OF PUPILS	19,212[1]
PRIMARY OUTCOME MEASURE AND SOURCE	Oral language skills (LanguageScreen standardised score)
SECONDARY OUTCOME MEASURE AND SOURCE	N/A

Study Plan version history

VERSION	DATE	REASON FOR REVISION
1.0 [original]	November 2022	

[1] Of the 548 schools included in the final pupil-level dataset only 434 schools completed the treatment status (NELI) indicator for their pupils, and therefore 114 schools cannot be included in the analysis. Of the 19,212 pupils, 15,573 remain after removing these schools.

Table of contents

Background and study rationale

The COVID-19 pandemic has disrupted oral language and literacy skills development in early childhood (Francis, 2022), with potential cascading effects on educational attainment and employability into adulthood (Feinstein & Duckworth, 2006). This disruption has also widened the existing gap in language skills and vocabulary between children from lower and higher income families, thereby disproportionately burdening pupils from disadvantaged backgrounds and/or with English as an Additional Language (EAL). Recognising the evidence in support of early interventions to narrow this gap, the Department for Education (DfE) has invested £17 million in the Nuffield Early Language Intervention (NELI) (DfE, 2021). In wave 1, rolled out in the 2020/21 academic year, NELI was offered at no cost to ~6,500 state-funded primary schools. Access to the programme was expanded to a further ~4,000 primary schools at no cost to the schools in wave 2 in the 2021-22 academic year. An independently commissioned implementation and process evaluation (IPE) of waves 1 and 2 of the NELI programme is currently underway.[2]

An EEF-funded individually randomised controlled efficacy trial conducted in 34 early years settings in 2016 demonstrated that two versions of the NELI programme – a 30-week programme starting at the end of nursery and continuing in the reception year and a 20-week programme offered in reception only – had a positive impact on the language skills of pupils. Children receiving the more expensive 30-week version experienced the equivalent of about four months of additional progress, whereas those experiencing the 20-week version experienced about two months of additional progress (Sibieta et al., 2016; Fricke et al., 2013). These positive findings were replicated in a larger EEF-funded cluster randomised controlled effectiveness trial in 193 schools where only the 20-week version of the programme was tested (Dimova et al. 2020; West et al., 2021). Children who received the NELI programme made, on average, an additional three months' progress in language skills compared to children who did not receive NELI. Children with EAL who received NELI also made the equivalent of three additional months' progress in language skills compared to EAL children who did not receive NELI.

While NELI has demonstrated a positive impact in trials, its impact *when delivered at scale* is not known. The aim of this scale up evaluation therefore is to assess the impact of NELI in 'real world' conditions. Contingent on the properties of the final pupil-level dataset, the evaluation will use a Regression Discontinuity (RD) design to assess the impact of NELI. The implementation of wave 2 of the NELI programme implies a data structure which supports this approach. At baseline, participating schools were asked to assess the language skills of *all* reception pupils using the LanguageScreen digital application and assign the 3 to 6 pupils with the lowest LanguageScreen standardised score to the NELI programme group. Effectively this would create a treatment assignment rule based on this criterion alone, but early findings from the wave 1 and 2 IPEs suggest that other selection criteria (besides the LanguageScreen standard score) were also used by teachers (see 'Selection Mechanism' section below). We discuss this in more detail herein and explain why a 'fuzzy' RD is therefore likely to be a suitable quasi-experimental design.

[2] See https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/nuffield-early-language-intervention-scale-up

Intervention

This is a scale-up impact evaluation of the Nuffield Early Language Intervention (NELI), developed by a team of researchers originally based at the University of York (see below for current affiliations) and funded by the Nuffield Foundation. The intervention aims to improve the oral language ability of reception pupils with relatively poor spoken language skills by focusing on their vocabulary, listening and narrative skills. The NELI logic model is described in detail in the study plan for the evaluation of the first year of NELI scale-up delivery (Dimova et al., 2021).

Aspect	Description
Name	Nuffield Early Language Intervention (NELI)
Why (Theory/Rationale)	The COVID-19 pandemic has had a serious impact on the development of oral language skills of children entering reception. Research undertaken by the University of York, the Education Policy Institute and the National Institute of Economic and Social Research suggests that 96% of schools reported being 'very concerned' or 'quite concerned' about their pupils' language and communication skills due to the pandemic (Bowyer-Crane et al., 2021). Strong oral language skills in children lay the foundation for literacy and numeracy (Duff et al., 2015; Law et al., 2013; Roulstone et al., 2011), essential for subsequent academic attainment and eventually employability in adulthood (Feinstein and Duckworth, 2006). The link between the development of language and literacy skills is well established (Whitehurst and Lonigan, 1998; Scarborough et al., 2009). It is also widely accepted that the process of learning begins early in the pre-school period and that support for those with poor language skills should be provided as early as possible (Whitehurst and Lonigan, 1998; Scarborough et al., 2009; Hulme and Snowling, 2013). There is also a policy imperative to intervene early as children from disadvantaged backgrounds are disproportionately more likely to suffer from language difficulties as they enter school (Fernald et al., 2013). Such disparities emerge early and tend to widen throughout the school experience, thus hampering efforts to improve social mobility. A significant body of evidence suggests that NELI has a positive impact on the language skills of children receiving the programme (Sibieta et al., 2016; Dimova et al., 2020; Bowyer-Crane et al., 2008; Fricke, et al., 2013; 2017; West et al., 2021). The programme has also been shown to be effective in improving the language skills of children with EAL, a group thought to be impacted by COVID-19 due to the reduced opportunity to hear and speak in English.
Who (Recipients)	Reception pupils (4 – 5 years old) with poor oral language skills in state-funded primary schools that have signed up to receive the NELI programme. Recruitment was first-come first-served with particular targeting of schools with a higher proportion of Free School Meal (FSM) eligible pupils, schools in local authorities in the bottom third of Ofsted ratings and schools in Opportunity areas. Participating schools were advised to screen all reception pupils using LanguageScreen,

	a short tablet-based standardised assessment. A report of children's scores could be downloaded and used to identify pupils in greatest need of targeted support. ~4,300 primary schools in England have signed up for wave 2 of NELI. Assuming 26.6 pupils per class, 1.5 forms per school and 3 to 6 NELI pupils per class, we estimate that 171,570 reception pupils will be screened and between 19,350 and 38,700 pupils will receive the NELI programme. The sample size for the scale-up impact evaluation is 548 schools and 19,212 pupils, although only 434 schools completed the treatment status indicator for their pupils meaning that data for 114 schools will need to be excluded from the analysis.
What (Materials)	Participating schools receive one NELI 'kit' per reception class. The kit consists of materials published by Oxford University Press and includes two teacher handbooks with detailed lesson plans, two sets of A4 flashcards and one Ted puppet to support session delivery.
What (Procedures, activities and/or processes)	**Training Model**: An online training programme developed by the University of Oxford was offered to nominated staff members (typically a Teaching Assistant (TA), reception class teacher and the school's NELI Lead) at participating schools. The training was hosted by OxEd and Assessment, Ltd on the FutureLearn platform. The core training content to be completed by the staff delivering the programme (usually the TA) consists of three linked online courses expected to take 10 – 12 hours to complete in a self-paced manner; two courses are to be completed before NELI delivery begins and the third short course can be completed halfway through programme delivery. Course 1 covers the fundamentals of understanding and supporting language development and an overview of the NELI programme. Course 2 covers the details of delivering the NELI programme in schools, including the screening process to identify eligible pupils and course 3 covers training on teaching letter sounds and phonological awareness. The reception class teacher is only expected to complete the first of the three linked courses although they will have access to the remaining two courses should they wish to complete the training. The NELI Lead (if different from the reception class teacher) also has the option to access the training if they would like to familiarise themselves with the programme. Staff members also receive ongoing online support throughout the course of the programme via the NELI Delivery Support Hub. All participating schools also received access to the LanguageScreen assessment tool. LanguageScreen has four subtests: expressive vocabulary, receptive vocabulary, listening comprehension and sentence repetition and takes about 10 minutes per pupil to administer. Schools were supported with identifying the appropriate pupils to receive the programme, usually 3 – 6 pupils with the weakest oral language skills, as assessed using LanguageScreen.
Who (Intervention providers/implementers)	NELI was developed by researchers led by Professors Charles Hulme and Maggie Snowling (now at the University of Oxford), including Silke Fricke and Claudine Bowyer-Crane, now at the University of Sheffield). It was funded by the Nuffield Foundation. The delivery of wave 2 of NELI is managed by a consortium of delivery partners led by the Nuffield Foundation

	Education Ltd (a special purpose vehicle of the Nuffield Foundation): • OxEd and Assessment – intervention developers; developers of LanguageScreen digital application; developers of the online training and delivery support model, hosted by them on the FutureLearn platform, providers of all support for intervention delivery (chat, email and in-person support for schools for both screening and programme delivery) and of weekly reporting to all delivery partners. Delivery support included employing and supervising the mentor team for the training and delivery support hub. This team included several members contracted via Elklan. • Oxford University Press – publish and distribute all materials that form part of the NELI kit provided to participating schools; provided targeted outreach during school recruitment Recruitment of schools for the scale-up impact evaluation was conducted by NFER.
How (Mode of delivery)	The TA or early years educator delivering the programme is expected to conduct 3 x 30-minute small group sessions consisting of 3 – 6 eligible pupils and 2 x 15 minute 1:1 support sessions per week for each targeted child. While NELI sessions are delivered during normal classroom hours, pupils selected to receive the programme are taken out of classes. School staff are responsible for identifying classes from which pupils were to be taken out to participate in NELI sessions. The group and individual sessions are to be scheduled on different days.
Where (Location of the intervention)	There were no geographical limitations on recruitment of schools for wave 2 of NELI. Within each participating school, it is recommended that the programme be delivered in a quiet area such as a classroom, library, staff room or dining room. Training was delivered online and could be completed at a location of the trainee's choice.
When and how much (Duration and dosage of intervention)	Recruitment of wave 2 schools took place from May to end of October 2021. Online training of staff was completed in the autumn term (September – December) of 2021. Following completion of training, intervention delivery was scheduled to begin in schools in January 2022 and to be completed by end of the summer term of 2022. The 20-week intervention is divided into two blocks of 10 weeks each. In the first 10 weeks, the intervention focuses on vocabulary, active listening and narrative skills and in the final 10 weeks also includes a small additional session element (3 minutes only) focusing on letter sounds and phonological awareness as foundations of early literacy skills. There is no minimum recommended dosage and participating schools are encouraged to deliver all 20 weeks of the intervention.
Tailoring (Adaptation of the intervention)	Tailoring is built into the programme through the bi-weekly 1:1 sessions; staff are encouraged to refer to their notes and observations and to tailor the 1:1 individual sessions to the needs of the pupils.

Costs	There is no cost to schools to receive the NELI programme; DfE funding covers all programme costs for participating schools. Schools participating in the impact evaluation will receive up to £250 for completing the following: (1) endpoint LanguageScreen assessments for the vast majority of reception pupils who received a baseline LanguageScreen assessment, (2) NELI indicator data marking which pupils received NELI and (3) final delivery survey administered by the Nuffield Foundation

About the evaluation

This study plan describes the NELI scale-up impact evaluation only, the aim of which is to estimate the impact of NELI when delivered at scale as part of a national roll-out. A separate implementation and process evaluation (IPE) is currently underway (Disley et al., forthcoming).

Impact evaluation

The impact evaluation is being undertaken to understand the impact of the NELI when delivered at scale on children's early language outcomes.

Research questions

The research questions all focus on one primary outcome, pupils' oral language outcomes as measured by the LanguageScreen standardised score post-intervention.

- **RQ1**. What is the impact of NELI when delivered at scale on pupils' oral language outcomes, as measured by LanguageScreen?
- **RQ2**. What is the impact of NELI when delivered at scale on FSM (everFSM) pupils' oral language outcomes, as measured by LanguageScreen?
- **RQ3**. What is the impact of NELI when delivered at scale on EAL pupils' oral language outcomes, as measured by LanguageScreen?
- **RQ4**. How does the impact of NELI on pupils' oral language outcomes vary by dosage?
- **RQ5**. How does the impact of NELI on pupils' oral language outcomes vary by training fidelity?

Design overview

Table 1: Design

Design	**Regression Discontinuity**
Unit of analysis (school, pupils)	Pupils

Number of Units to be included in analysis (Intervention, Comparison)		11,351 (2,755 intervention, 8,596 comparison)[3]
Primary outcome	variable	LanguageScreen standardised score (endline)
	measure (instrument, scale, source)	LanguageScreen
Secondary outcome(s)	variable(s)	N/A
	measure(s) (instrument, scale, source)	N/A
Baseline for primary outcome	variable	LanguageScreen standardised score (baseline)
	measure (instrument, scale, source)	LanguageScreen
Baseline for secondary outcome	variable	N/A
	measure (instrument, scale, source)	N/A

The impact evaluation is based on a quasi-experimental design which makes use of the large volume of secondary data collected by the LanguageScreen assessment system during the course of the scale-up. Schools were asked to assess all children in their Reception classes before commencing delivery of NELI, and again in the 2022 summer term, after around 20 weeks since beginning delivery to children. A regression discontinuity is implied by the intervention scale-up's recommendation that teachers effectively rank children within each class (by LanguageScreen baseline assessment scores) in order to determine treatment assignment, such that the lowest 3 to 6 scoring pupils within each class are designated to receive NELI.

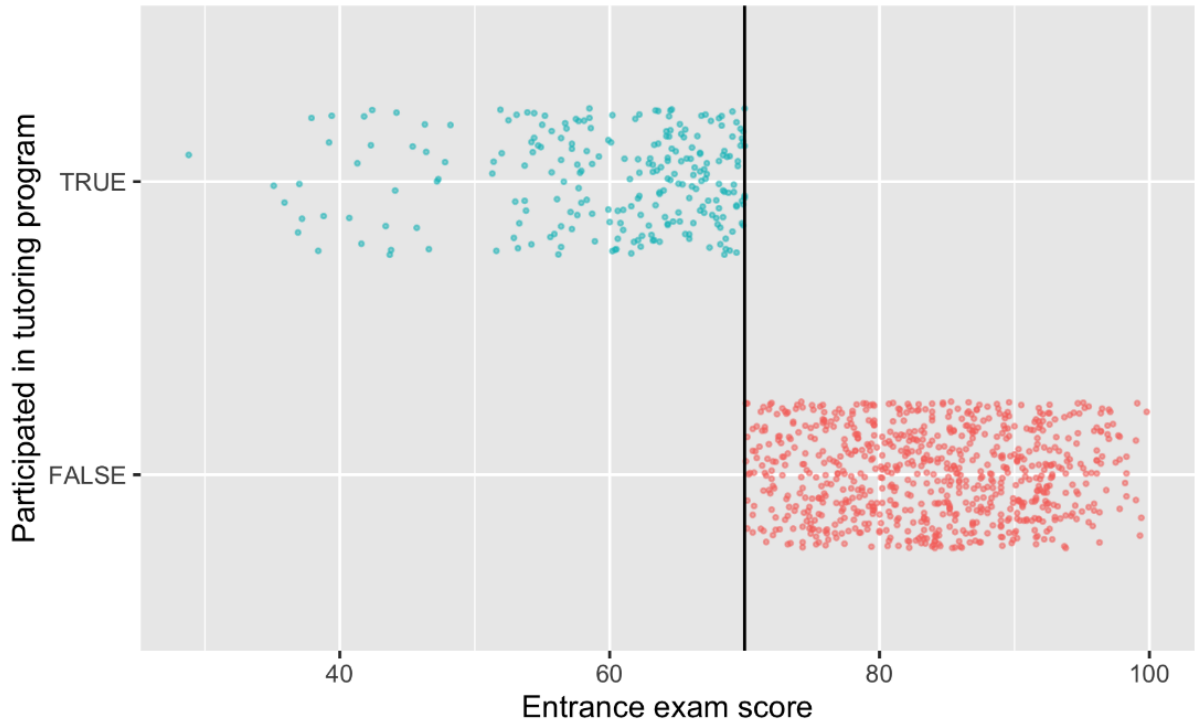
In general terms, all units in an RD design receive a score with treatment being assigned to all units above (or below) a known point on the score range. Figure 1 below (Heiss, 2022) illustrates this using simulated data. In this case an entrance exam score of 70 or below determines participation in the tutoring programme. Entrance exam score is the **running variable**, and 70 is the **cutoff** or **cut score**. This is an example of a **Sharp RD** (SRD) where the running variable uniquely and absolutely determines treatment assignment.

In the NELI evaluation the LanguageScreen standardised score (at baseline) is the running variable, and the cutoff is partially determined by the ranking process (see paragraph immediately following). As this cutoff is based on ranking rather than any absolute value, the design accommodates the fact that the cutoff value varies by class. This is an example of **noncumulative multiple cutoffs** (Cattaneo et al., 2016) whereby units in different groups

[3] These numbers indicate pupils for whom LanguageScreen endline assessments have been completed and who also have their NELI indicator status completed. Of the 19,212 pupils in the final pupil-level dataset, 15,573 remained after schools (n=114) which had not completed the NELI status indicator of their pupils were removed. Of these, 4,222 were not re-tested at endline, leaving 11,351 for analysis.

(e.g. schools) receive a rating on a running variable, but the cutoff varies by group. To address this, the cutoff for each class is subtracted from class members' individual scores, thus re-centering the running variable and normalising the cutoff at zero. This allows for the data from multiple classes to be pooled and analysed together (ibid.).

Figure 1. Sharp RD (simulated data)



The NELI Wave 2 IPE found that the recommendation to assign treatment based on ranking was not followed uniformly by all staff involved in deciding which children get NELI (see 'Selection Mechanism' below). This implies that within each class, ranking (and therefore the cutoff) was not absolutely deterministic of treatment and that treatment assignment was therefore also based on other criteria. Where this is the case, a different type of RD design is implied. A **Fuzzy RD** (FRD) is one in which units either side of the cutoff may be treated or untreated. In this case, the cutoff implies a probability of receiving treatment rather than determining it absolutely. Other factors may also play a part in the allocation of some units to treatment, and these maybe either known or unknown to an evaluator. Furthermore, they may be unobservable in the evaluation data (Van der Klaauw, 1997). This does not present an insurmountable problem if the probability of receiving treatment based on the running variable is higher than a certain threshold (point 4 in 'preliminary analysis' section). Figure 2 below illustrates this using initial pupil-level data received by the evaluators in August 2022. It shows pupils who were selected for NELI, the majority of whom scored below the cutoff, and pupils who were untreated, the majority of whom scored above the cutoff.

As a regression discontinuity design is a property of the data structure (rather than being a purely analytical choice), NFER undertook **initial preliminary analysis** to confirm the viability of the proposed design. This was based on an initial pupil-level dataset (received by NFER from OxEd in August 2022). Details of the initial preliminary analysis can be found below, and these data confirmed that the fuzzy RD design would be feasible. Before proceeding with analyses to estimate impacts of the intervention, NFER will re-run the preliminary analysis on the final dataset.

Figure 2. Fuzzy RD (NELI initial pupil-level data)

The initial preliminary analysis was performed using a subset of the initial pupil-level data set. After exclusion of duplicate test IDs and classes with all pupils or no pupils in receipt of the intervention, we explored a data set of 14,195 pupils from 584 classes in 402 schools. A cutoff for each class was determined using the method detailed in 'selection of the comparison group and identification assumptions'. A histogram of the cutoffs can be found in Appendix A (figure 5). These cutoffs were used to adjust the pre-intervention LanguageScreen scores so that a cutoff of 0 was appropriate for the whole data set.

Comparing the intervention received to whether the pre-intervention LanguageScreen score was above or below zero, we determined that the percentage of 'non-compliers' was 6% (Figure 2). The What Works Clearinghouse Handbook (2022) indicates that study authors must run the first-stage regression of the participation indicator on the forcing variable, and the indicator for being above or below the cutoff, and provide the F statistics from this regression. In the preliminary analysis, the F statistic for the forcing variable was 85.85 and the F statistic for the indicator for being above or below the cut-off was 164.89. The low percentage of non-compliers and high F statistics suggest that the fuzziness in our data set is not extreme, so a FRD analysis is acceptable.

Graphical illustration of the probability of intervention receipt against adjusted pre intervention LanguageScreen score (Appendix A, figure 6) also supports this approach. Mass points (see also 'preliminary analysis' section later for further explanation) were investigated in the preliminary data and we found that the 10,703 observations (with non-missing endline LanguageScreen scores) took 191 unique values. The number of mass points is considered sufficiently large for a continuity-based approach.

Participants

Participating institutions are English schools with one or more Reception classes in the 2021/22 academic year which are taking part in NELI Wave 2. These schools had previously

agreed to being contacted about the evaluation and were subsequently contacted by NFER to ask whether they would be willing to take part. Pupils eligible for inclusion in the evaluation are those in participating schools who have had a LanguageScreen baseline assessment and who have not been withdrawn from the study by their parents[4]. Treatment and comparison groups are formed by a number of criteria, including baseline assessments of language ability (LanguageScreen standardised score). These are discussed in detail in following sections ('selection mechanism' and 'selection of the comparison group and identification assumptions'). The final treatment sample was defined as at the end of September 2022 (although outcomes were measured in the summer term of 2022). Additional data collection took place in September on the NELI treatment indicator variable as some schools had not completed this for their pupils; all schools were required to enter this into the LanguageScreen application as pupil receipt of NELI was not automatically recorded when staff undertook baseline testing of pupils.

Sample size calculations

As it was necessary to recruit schools to the impact evaluation from the approximately 4,300 schools taking part in NELI Wave 2, NFER carried out initial sample size calculations for a prospective RD study without data. These were based on Deke and Dragoset (2012) and McKenzie (2022). Using the former we estimated the number of schools required to achieve different Minimal Detectable Effect Sizes (MDESs). These estimates were based on the assumption of 26.6 pupils per class with an average of 1.5 classes per school and a pre-/ post-test correlation of 0.5[5]. They were also based on three different Regression Discontinuity Design Effects, i.e. the number of times to multiply the equivalent RCT sample size for an RD design (9, 14 or 17). Using McKenzie's (2022) guidance this number was then adjusted based on three different levels of assumed fuzziness under the anticipated Fuzzy RD design (e.g. a probability of 0.5 instead of 1 in the chance of being treated below the cutoff – repeated for 0.7 and 0.9). These methods assume some properties about the data but do not include others which are unknown in the absence of data. For example we did not know: i) the distribution of scores on the running variable (LanguageScreen standard score at baseline); ii) the location of the cutoff; iii) the proportion of treated and control pupils in the sample (although for the latter we assumed 23:77 based on 5-6 pupils per class receiving NELI). Following discussion with EEF and OxEd we selected a number of schools to recruit based on a target MDES of 0.2 (see Appendix B).

Once initial pupil-level data had been collected, we were able to use this data to inform a second sample size calculation. This utilised the rdsampsi function in the *rdpower* R package (Cattaneo et al., 2019). Data variability and fuzziness was taken from the initial data. Two mean square error bandwidth selectors (above and below cut off) were selected as the bandwidth selection procedure. Mass points (see 'preliminary analysis' section later) were addressed by requiring that initial bandwidths contain at least 10 unique values. Pre-intervention LanguageScreen score was included as a baseline covariate. Due to the normalisation of outcome scores to a class level cutoff, we expect no variability in mean scores

[4] 28 children were withdrawn from the study.

[5] 26.6 pupils per class (https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics); 1.5 forms per school (NELI scale-up Y2 impact evaluation ITT - Feb 2022); Pre-/post-test correlation as per Deke and Dragoset (2012).

across classes so a multilevel model approach was not implemented. While the distribution of residuals in the final dataset is not yet known, the preliminary data suggested that heteroskedasticity might be expected. Therefore, a heteroskedasticity-robust plug-in residuals variance estimator was selected as the variance-covariance estimator. Sample size calculations were at a pupil level and class numbers in Table 2 are based on average cluster size.

Table 2: Sample size calculations

		Initial calculation (without data)	Calculation with initial pupil-level data (n=10,703)	
		OVERALL	OVERALL	FSM
Minimum Detectable Effect Size (MDES)		0.2	0.2	0.46
Pre-test/ post-test correlations	level 1 (pupil)	0.5	0.86	0.86
	level 2 (class)	-	-	-
	level 3 (school)	-	-	-
Intracluster correlations (ICCs)	level 2 (class)	-	-	-
	level 3 (school)	-	-	-
Alpha		0.05	0.05	0.05
Power		0.8	0.8	0.8
One-sided or two-sided?		Two-sided	Two-sided	Two-sided
Average cluster size		27	26	5*
Number of classes	Intervention	-	-	-
	comparison	-	-	-
	Total	702	509	509
Number of pupils	Intervention	3,848	3,202	602
	comparison	14,833	10,034	1,888
	total	18,681	13,236	2,490

* Assuming 18.81% of pupils are eligible for FSM (everFSM). Based on state funded primary school reception pupils in 2021/22 from https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics.

After cleaning, the initial data contained 10,703 pupils with pre- and post-intervention assessment data (out of 14,272 pupils, of which 3,569 pupils did not have outcome data). This sample size would give a MDES of 0.22. Further data collection (of the NELI indicator status) which took place in schools in September 2022, marginally increased the sample size to 11,351 pupils with both pre- and post-intervention assessment data. This should result in a reduced MDES (lower than 0.22), when the sample size is calculated during preliminary data analysis.

Outcome measures and other data

Baseline measures

We will use the LanguageScreen standardised score from the baseline screening assessments (completed before the intervention was delivered to pupils) as our baseline measure.

Primary outcome

Our primary outcome will be the LanguageScreen standardised score from endline assessments. These were completed in the Summer of 2022, around 20 weeks after the intervention started being delivered to pupils. The LanguageScreen application is a screening assessment administered by TAs which takes around 10 minutes to complete and is scored automatically. It comprises four subtests (West et al. (2022)):

1. Expressive Vocabulary (naming 24 pictures)
2. Receptive Vocabulary (matching each of 31 spoken words to one of four pictures)
3. Sentence Repetition (repeating each of 12 sentences verbatim)
4. Listening Comprehension (answering 12 questions on spoken stories that tap literal and inferential comprehension)

The LanguageScreen score is derived from the four subtests as a latent variable (West et al., 2021), with the standardised score being based on a sample of 348,944 children which was used for standardisation (G West, email 2022)[6]. West et al. (2021) reported that LanguageScreen reliability was found to be high (pre-test screening Cronbach's $\alpha = 0.84$) in the effectiveness RCT with good concurrent validity, whilst the standardisation process described herein found the LanguageScreen total score to have excellent reliability (Person Separation Reliability = .94).

Secondary outcomes

None

Other data

Details of variables to be used can be found in Appendix C. OxEd will send pupil-level data from the LanguageScreen system to the Department for Education for matching with National Pupil Database pupil data, school-level characteristics and school-level dosage and fidelity data. The matched dataset will be made available to the evaluation team for analysis using the Secure Research Service.

[6] Standardisation was categorised into 6 month age bands (48 – 53 months n = 55,306; 54 – 59 months n = 157,642; 60 – 65 months n = 115,400). Data was analysed using Rasch modelling and provided a good fit to the model (RMSEA = .03, SRMSR = .09).

- LanguageScreen data will include the baseline and primary outcome measures described above, school and class identifiers, variables which describe pupils (pseudonymised) and their characteristics (e.g. date-of-birth, EAL)
- NPD data will include variables for each pupil included in the LanguageScreen data, contributing further characteristics (e.g. Free School Meals, Special Educational Needs) which will be used to determine balance across treated and untreated groups and for covariate adjustment.
- School-level characteristics derived from DfE's publicly available Get Information about Schools dataset (and where necessary supplemented by data from NFER's Longitudinal Schools Database; for example, where data are missing from the former, or to resolve URN matching issues). School-level characteristics will be used for our descriptive analysis of the sample.
- School-level dosage and fidelity data derived from Nuffield Foundation Ltd's delivery surveys and OxEd's school staff training. Dosage will be derived from the survey item 'What was the session number of the last group session delivered?' (a maximum of 60, assuming 20 weeks of programme delivery and a recommendation of 3 sessions per week). Training fidelity will be calculated by comparing the number of active Teaching Assistant (TA) trainees for each school with the number of TAs per school who have completed training courses (for each of three training courses).

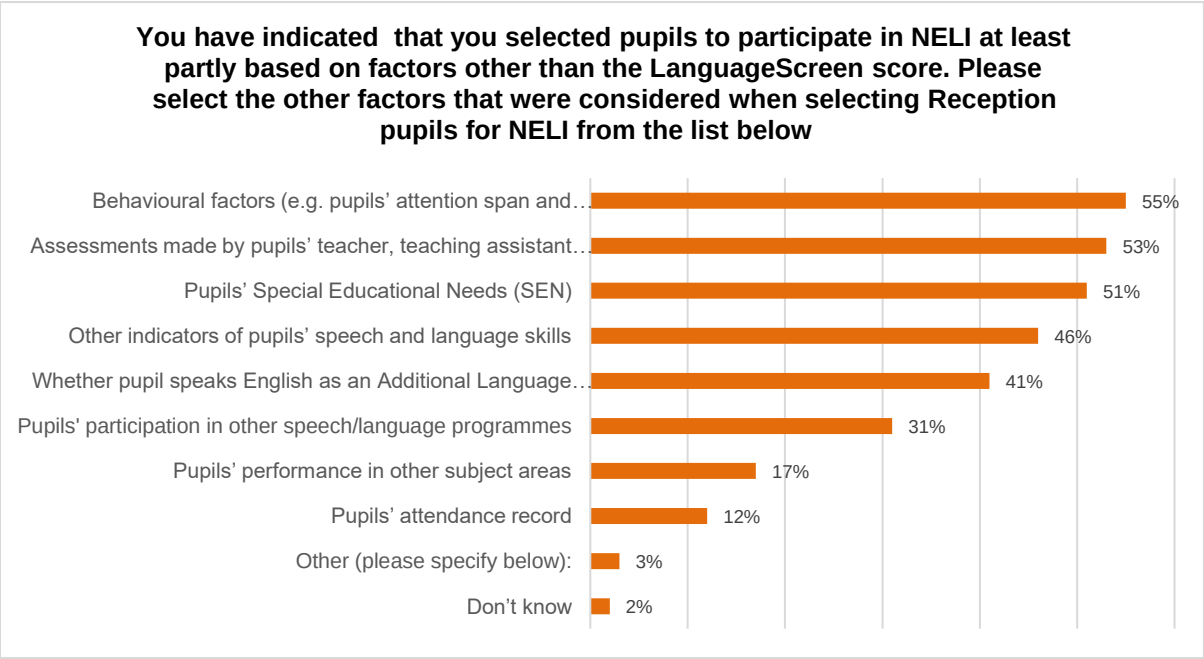
Selection Mechanism

All children are thought to be able to benefit from NELI[7] and there are no specific criteria determining ineligibility for the intervention. In order to select pupils to take part in NELI, it was suggested by the NELI developers that schools should use LanguageScreen to undertake a baseline assessment of pupils' oral language abilities, using this data to rank pupils within each class, selecting the pupils with the lowest scores for NELI. This would therefore imply a single criterion for selection into treatment using a numerical variable. Furthermore, variation in the absolute score used to determine treatment is expected across classes and schools, given the use of ranking and the relative difference in abilities across schools (i.e. schools with pupils who on average are of higher ability will have a higher cutscore than those with lower than average ability children). Children receiving NELI will in most cases have LanguageScreen scores below average, with schools effectively ranking pupils to select them for the intervention.

The RAND IPE found that only 33% of schools (of 524 schools surveyed in December 2021) selected pupils based on LanguageScreen scores alone, whilst 66% selected pupils based on other factors in addition to LanguageScreen. 121 schools provided further information about these (figure 3) (Disley et al., forthcoming).

Figure 3. IPE findings on factors influencing selection of pupils for NELI

[7] https://www.teachneli.org/faqs/



Therefore, in the majority of schools the selection mechanism was not based on a single criterion but a range of factors, with schools using multiple criteria in addition to LanguageScreen baseline scores. For individual pupils these additional criteria are not known by the impact evaluation team, and as the graph above indicates, some are not recorded in administrative data.

The precise selection mechanism therefore varies across classes, suggesting that some use one criterion (LanguageScreen scores) whilst others also use one or more other criteria. The implication for RD analysis is that a Fuzzy RD (FRD) is appropriate, given that the running variable (LanguageScreen baseline score) does not uniquely determine treatment in all schools. Given also that cutoff scores are not absolute and vary by school it is likely that they will cluster around a similar point in the score distribution (corresponding to the boundary between the LanguageScreen assessment descriptions 'below average' and 'low average').

Selection of the comparison group and identification assumptions

As the NELI intervention represents a binary treatment (pupils are either selected for NELI or they are not), the selection mechanism above therefore creates the treatment and comparison groups for analysis. The pool of comparison pupils is defined as those individuals not in receipt of the NELI intervention. For a straightforward sharp RD, these would be the individuals with values above the cutoff on our running variable (baseline LanguageScreen standard score). However, in practice there is evidence of non-compliance to this selection mechanism (i.e. fuzziness) due to other criteria also being used by school staff. There is also selection at a class level implying multiple cutoffs which are not defined. Our proposed methodology for addressing these issues is to select the cutoff for each class that minimises the number of non-compliant pupils, normalising the cutoff to be zero across all classes by subtracting the cutoff from the scores, and allowing for fuzziness in the analytical approach. The method for selection of a cutoff for each class will be:

1. Set the cutoff to be the minimum baseline LanguageScreen score for the class and calculate the proportion of pupils correctly classified as intervention or comparison.

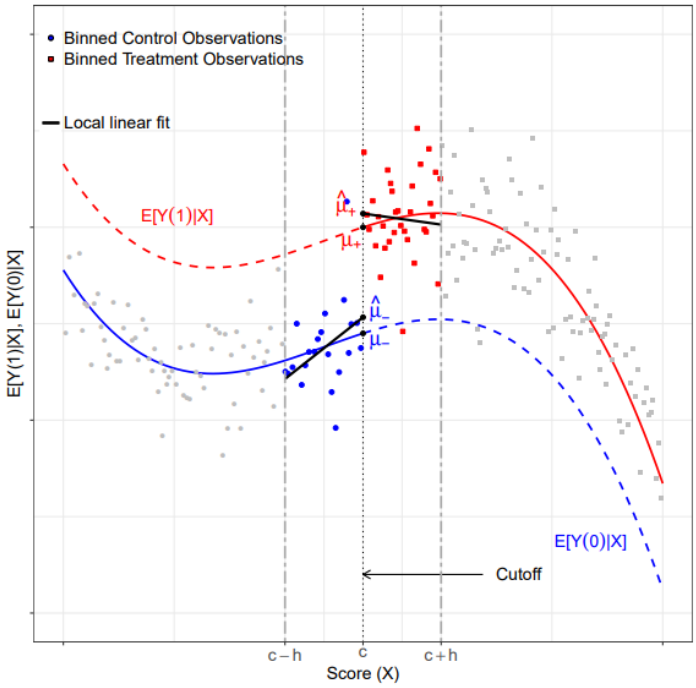
2. Repeat using each integer between the minimum and maximum baseline LanguageScreen score for the class as the cutoff.
3. Take the median of the assessed cutoffs which produce the maximum classification accuracy.

Analysis of Regression Discontinuity designs

RD analysis is undertaken within one of two frameworks, one founded on continuity-based assumptions and the other based on local randomisation assumptions (Cattaneo et al., 2019), with both relying "on the assumption that units that receive very similar score values on opposite sides of the cutoff are comparable to each other in all relevant aspects, except for their treatment status … The main distinction between these frameworks is how the idea of comparability is formalized: in the continuity-based framework, comparability is conceptualized as continuity of average (or some other feature of) potential outcomes near the cutoff, while in the local randomisation framework, comparability is conceptualized as conditions that mimic a randomised experiment in a neighborhood around the cutoff." (ibid, p.5).

Both approaches therefore imply a comparison between units close to either side of the cutoff. To select these units a **bandwidth** is required, illustrated by figure 4 below. In this example the cutscore is *c*, and the bandwidth is indicated by *c*-*h* and *c*+*h*, with the same bandwidth (*h*) being used on either side. Estimation occurs within this window, using the units contained therein. The choice of bandwidth therefore has important implications for impact estimates and we discuss below our approach to **bandwidth selection** and sensitivity analysis. We also discuss a range of checks, one of which is for **mass points** on the running variable. This is the case where the running variable is not continuous, such that many observations may have the same value. We know this to be true of the LanguageScreen standardised score and discuss this further below in 'preliminary analysis'.

Figure 4. RD estimation (continuity-based approach; Cattaneo et al., 2019, p. 43)



Preliminary analysis

The evaluation team have undertaken initial preliminary analysis using initial pupil-level data available to us in August 2022. This was to produce more accurate sample size calculations than those without data allowed for, and to test whether an RD design would be viable. From this analysis we conclude that the design is viable, but in order to ultimately confirm this we will undertake **comprehensive preliminary analysis** using the **final pupil-level data**. This will be in advance of any primary and exploratory analyses and will also serve to understand the final dataset in detail in the context of the proposed FRD design. This step will therefore help to shape final analytic choices. Specifically, this comprehensive preliminary analysis will enable testing of some of the RD identifying assumptions and modelling noncumulative multiple cutoffs in the data.

Preliminary analysis will include the following:

1. Describing the characteristics of the school sample, e.g. schools in the evaluation sample compared with all English primary schools (with Reception classes).
2. Analysis of LanguageScreen baseline scores to model cutoffs by class, including calculating pupil baseline scores relative to the class cutoff. This will follow the same method as outlined in ‘Selection of the comparison group and identification assumptions’, i.e. maximising classification accuracy
3. Graphical analysis to confirm that the proposed FRD design is valid, for example by checking for treatment assignment either side of the cutoff, and plotting the outcome against the cutoff to visually inspect for evidence of a discontinuity.
4. Estimating the probability of receiving treatment as a function of the LanguageScreen cutoff. As FRD can be understood as an instrumental variables model (with the cutoff functioning as an instrument which determines treatment probability) it is important to check that the instrument is not weak, i.e. that there is a substantial difference in treatment assignment either side of the cutoff. The What Works Clearinghouse Handbook (What Works Clearinghouse, 2022) indicates that study authors must run the first-stage regression of the participation indicator on the forcing variable, and the

indicator for being above or below the cutoff, and provide either the F statistic from this regression, and that “an F statistic of 16 will be used as the interim criterion for assessing instrument strength”.

5. Determining the number of mass points in the data (due to non-continuous data on the running variable), and whether this precludes adopting a continuity-based RD approach (as per Cattaneo et al., 2020, pp.60-62).
6. Power calculations based on final sample size. By using the final pupil-level dataset we will be able to estimate the MDES with much greater precision than our initial calculations without data allowed (see 'Sample size calculations’). The initial calculations were not able to incorporate some of the properties unique to the dataset, such as the number of observations on each side of the cutoff and the degree to which the cutoff determines treatment (i.e. fuzziness). We will use the rdpower package in R or Stata for these estimations, based on:
    a. outcome data (LanguageScreen endline assessment) for the final pupil sample
    b. the modelled cutoffs (by class) with pupil running variable scores relative to these
    c. known treatment assignment
    d. treatment probability inherent in the data
    e. covariates, including pre-test scores
    f. bandwidth selection using two different mean squared error (MSE) optimal bandwidth selectors (below and above the cutoff) for the RD treatment effect estimator (see note on bandwidth selection below)
    g. heteroskedasticity-robust plug-in residuals variance estimator without weights
7. Balance checks. We will investigate balance on characteristics such as gender, month of birth, FSM, EAL, and SEN. Some of these characteristics (e.g. SEN, EAL) may be criteria used by some teachers to select pupils for NELI and hence could potentially lead to imbalance either side of the cutoff. However this may be offset by those schools which are using LanguageScreen alone to select pupils, and the fact that the data are pooled across schools (which suggests that characteristics may be better balanced due to between-school variation). An analysis of the initial pupil-level data showed for example that EAL pupils are relatively well-balanced below the cutoff, although this test did not apply a bandwidth to select cases. To test for balance under the local polynomial approach which we intend to use for our primary analysis we will check for statistically significant differences in these characteristics either side of the cutoff. If these tests suggest an imbalance there are two options which may be suitable, depending on the nature and extent of the imbalance: i) proceed with primary analysis using the local randomisation approach (which we have already specified to test the robustness of our assumptions regarding the viability of the local polynomial approach), or ii) redefine the parameter of interest (Cattaneo et al., 2021), e.g. as a weighted average of treatment effects, although we note that this is not preferable due to the additional assumptions that this may imply and the fact that this is of less interest for the purpose of impact evaluation.
8. Missing data (see ‘Missing Data’ section below).

Primary analysis

We will proceed with impact analyses using the continuity-based approach to RD described by Cattaneo et al. (2019). Within this framework we will proceed with intention-to-treat analysis

of attainment outcomes using LanguageScreen standard scores at endline and given the FRD design will estimate a Local Average Treatment Effect. We will give careful consideration to our interpretation of estimands in the context of the final model used and underlying assumptions, particularly with regard to their generalisability; to pupils within the bandwidth but distant to the cutoff, to pupils outside the bandwidth, and to pupils outside the sample.

During preliminary analysis pupils' scores on the running variable (LanguageScreen standard score at baseline) will have already been calculated relative to a cutoff which has been normalised at zero. As such our primary analysis will address the noncumulative multiple cutoffs in the data (Cattaneo et al., 2016), and aligning them as such across classes and schools serves to mitigate clustering effects. Therefore we will not address this further in our analysis.

We will estimate impacts using a suite of R packages written specifically for RD analysis within the continuity-based approach[8]. These include:

- rdbwselect for data-driven bandwidth selection methods
- rdrobust for local polynomial point estimation and inference
- rdplot for graphical RD analysis

In essence this will require us to choose some of the parameters for the estimation, including bandwidth. To avoid the introduction of evaluator bias (e.g. testing different bandwidths until a preferred impact estimate is obtained), we will allow the appropriate bandwidth to be determined in a data-driven manner, but will conduct sensitivity tests using different bandwidths.

RD estimation under continuity based assumptions involves fitting two weighed least squares regressions, one each side of the cutoff and calculating a point estimate ($\hat{\mu}+ - \hat{\mu}-$; as illustrated by figure 4). Primary analysis will be carried out with three models. Cattaneo et al. (2021) note that researchers should always report unadjusted and covariate adjusted RD results and that the latter should not change the point estimate but may improve efficiency and reduce confidence intervals:

1. An unadjusted model. LanguageScreen standardised score (endline) regressed on pupil's score relative to cutoff, weighted for each observation, using cluster-robust standard errors.
2. A covariate adjusted model (pre-test): as model 1, including also LanguageScreen standardised score (baseline) as pre-test covariate.
3. A covariate adjusted model (pre-test plus other pupil-level covariates): as model 2, including also gender, month of birth, FSM status, EAL status, and SEN status.

Our second and third models will require us to carry out further balance checks using the pupils included in the analysis by the bandwidth selection. If covariates are balanced across these groups then we will interpret the covariate adjusted model under the same assumptions as Model 1, and we will report Model 3 as our headline finding. If however they are imbalanced then we will need to rely on different assumptions or redefine the impact estimate parameter.

[8] Further details can be found at https://rdpackages.github.io/

Inference

We will select an inference method which is congruent with the continuity-based (non-parametric) approach, such as robust bias correction (RBC), a method which can be simply implemented using the aforementioned packages and which has been found to perform well in a range of contexts and in simulation studies (Cattaneo and Titiunik, 2022.)

Robustness and sensitivity checks

We will test the robustness of our findings to assumptions implied by the continuity-based approach and the FRD design, and their sensitivity to analytical choices. These checks will include:

1. Density of the running variable, to check whether the number of observations just below the cutoff is greatly different to the number of observations just above the cutoff. If this were the case it would suggest some manipulation (i.e. that units could manipulate the score they receive on the running variable to influence treatment status), and would potentially undermine some of the assumptions on which RD is based. We will use the rddensity package which provides manipulation tests of density discontinuity based on local polynomial density estimation methods (Cattaneo et al., 2018).
2. Exclusion of observations near the cutoff. By running analyses excluding units very close to the cutoff we can understand the reliance by estimators on those units. This step is also relevant when there is evidence of manipulation (point 1 above).
3. Treatment effects at placebo cutoffs. We will check for treatment effects at cutoff values other than the actual cutoff (i.e. 0). RD analysis is based on the assumption that regression functions are continuous at points other than the cutoff, i.e. that there should be no discontinuities away from the cutoff.
4. Sensitivity to bandwidth choices. By varying bandwidth choice from that selected by MSE-optimal methods, we will effectively add pupils to, or remove them from, the area of bandwidth within which the estimations are made. Adding pupils by increasing the bandwidth has the effect of decreasing variance (and hence, confidence intervals) but increasing bias. We will investigate our findings under alternative bandwidth choices.
5. Estimations excluding classes where treatment assignment is fuzzy (i.e. running variable does not uniquely determine treatment). This will allow us to effectively analyse a subset of data as a sharp RD. In the initial preliminary analysis, 38% of classes had a sharp cut off.
6. Testing the alternative RD framework (i.e. the local randomisation approach, assuming that we proceed with the continuity-based approach for our analyses). Doing so tests some of the fundamental assumptions of the preferred approach, for example, the continuity of potential outcomes near the cutoff.

Our robustness checks may therefore indicate that an alternative approach to RD (than our suggested continuity-based approach) is warranted, and if so it may be necessary to rerun our analyses accordingly.

Further analyses

Subgroup analyses

We will explore treatment effect heterogeneity by repeating our primary analysis for FSM (using EVERFSM_6_P_[term][yy][9]) and EAL pupils. As these are relatively straightforward binary variables, we will do this using indicator variables for each category to subset the data, and then will proceed to estimate treatment effects for the different groups (e.g. FSM and non-FSM). This does not require any additional assumptions for estimation (i.e. additional to those which are required by the continuity-based approach).

Exploratory analyses

We will explore treatment effect heterogeneity by dosage (number of group sessions delivered) and training fidelity. This data is derived from school surveys (completed in July 2022) and school staff training data supplied by OxEd. Surveys were returned by 344 schools and we therefore consider this analysis to be exploratory as this is not the complete number of schools for whom we have pupil-level data, and also the fact that schools returning surveys may not be representative of the entire sample. We will produce descriptive statistics comparing those schools which returned a survey with all schools in the sample to determine this.

Both variables will be included separately in models which build on those of the primary analysis and we will check for balance at the cutoff in both the treated and untreated groups of pupils.

Effect size calculation

We will calculate effect sizes (Hedges g; Hedges, 1981) for primary, secondary and subgroup analyses. The RD model coefficients will be divided by the pooled variance. Confidence intervals for coefficients will be calculated using the Normal approximation.

Missing data

Based on our understanding of schools' use of LanguageScreen and their evaluation activity, we anticipate that the matched (with NPD) data will be complete with the exception of NELI indicator status and/or LanguageScreen endpoint assessment for a minority of pupils. Regarding the latter, we assume missingness to be determined by school and teacher decisions about whether to re-test pupils post-intervention, and in some instances which groups of pupil (intervention / non-intervention) to re-test. Furthermore, schools were encouraged to indicate in LanguageScreen which pupils were receiving NELI as a requirement of the evaluation, although not all schools ultimately completed this step.

We will specify the number of complete cases. To investigate our assumptions about missingness we will explore the nature and extent of any missing data. We will check for entire

[9] In line with EEF statistical analysis guidance. Due to age of pupils this is expected to be the same as FSMeligible_[term][yy]

schools which have missing data on NELI indicator and LanguageScreen endpoint assessment variables and we will compare their characteristics with schools with complete data. We will model missingness firstly at the school-level by means of a logistic regression where the outcome (for all pupils in the school) is either missing NELI indicator or missing LanguageScreen endpoint assessment. We will also model missing data at the pupil-level using two-level (pupil and school) logistic regressions with the same outcomes. We will include all pupil-level covariates which we will specify in our analysis (to include LanguageScreen baseline score, EAL, SEN etc.) in addition to school-level covariates.

We note the possibility of false discoveries in regression discontinuities should the missing data mechanism be related to the LanguageScreen endline assessment outcome (De la Cuesta and Imai, 2016). Therefore, having explored the pattern of missing data, we will proceed in line with the road map from EEF 2018 analysis guidance.

Ethics

The evaluation will be conducted in accordance with NFER's Code of Practice, available at NFER Code of Practice. All of NFER's projects abide by its Code of Practice, which is in line with the Codes of Practice from BERA (the British Educational Research Association), MRA (the Market Research Association) and SRA (the Social Research Association), among others. NFER is committed to the highest ethical standards in all of its activities and ethical considerations are embedded in its detailed quality assurance processes.

Each school's head teacher gave permission for the school's participation in the impact evaluation by signing a Memorandum of Understanding (MoU). Parents of all reception pupils in schools participating in the impact evaluation were sent a letter explaining the evaluation and given the opportunity to withdraw their child from data processing.

Data protection

Data Protection Statement and GDPR Compliance

NFER is registered with the Information Commissioner's Office for all of our research and other activities. We ensure that all projects comply with the seven principles of data protection legislation (UK General Data Protection Regulation (GDPR) and the Data Protection Act 2018).

Our Head of Data Security oversees NFER's compliance with data protection legislation and the ISO/IEC framework for information security, ensuring that policies and practice are both legally compliant and good practice. NFER is ISO/IEC 27001 certified (GB17/872763), and we hold Cyber Essentials Plus (IASEM-CEP-004922). We maintain a full Information Security Management Strategy (ISMS) including a Data Security Policy with which all staff are required to comply.

To carry out this evaluation, it will be necessary to use and share personal data about pupils (both those who received the NELI programme and those who did not) as well as key staff members including the school's head teacher, NELI Lead and bursar or equivalent. Additionally, we will also receive special category data (i.e. ethnic origin and special education needs) for pupils matched from the NPD.

NFER will receive pseudonymised pupil data from OxEd and Assessment for the purpose of preliminary analyses to assess the viability of the RD approach and to identify schools that have met the requirements for incentive payments. Any data sharing between NFER and OxEd and Assessment will be carried out via a secure portal.

For further information, please see the NELI impact evaluation privacy notice, available at: https://www.nfer.ac.uk/media/4945/nelv_school_information_sheet.pdf

Legal bases

The legal basis for processing the personal data accessed and generated by this trial is covered by GDPR Article 6 (1) (f) which states that;

'processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party except where such interest are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of the personal data'.

We have carried out a legitimate interest assessment, which demonstrates that the evaluation fulfils the NFER's core business purposes (undertaking research, evaluation and information activities). It has broader societal benefits and will contribute to improving the lives of learners by providing evidence for about the most effective ways of enabling pupils with poor oral language skills to catch-up with their peers. The evaluation cannot be done without processing personal data but processing does not override the data subject's interests.

The legal basis for processing pupils' special personal data is covered by GDPR Article 9 (2) (j) which states that:

'processing is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with Article 89(1) (as supplemented by section 19 of the 2018 Act) based on domestic law which shall be proportionate to the aim

pursued, respect the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject'.

We do not believe this processing will cause damage or distress to the data subjects.

Linking to NPD and Use of the Secure Research Service (SRS)

OxEd and Assessment will securely submit pupil-level data including pupil identifiers to the National Pupil Database (NPD) team to be matched to the pupil data held in the NPD. NFER will also securely submit school-level data to the NPD team. The NPD team will create a composite dataset that includes pupil-level and school-level data and the matched pupil data from the NPD. The team will also replace pupil identifiers with the Pupil Matching Reference and remove any direct school identifiers. NFER can only access these data within the SRS and any outputs will be checked to ensure that no pupils can be identified.

The project meets the Office for National Statistics "five safes" in the following ways:

- Safe people: all researchers accessing the project's data via the SRS are Accredited Researchers and hold a 'basic disclosure' certificate that is no more than 2 years old
- Safe projects: the project meets the conditions for accessing personal level data. A full request to the NPD team will be submitted, outlining the appropriate and ethical use of the data, and the public benefit of the research and its broader societal benefits. The research team and the EEF are committed to publishing the results of the study.
- Safe settings: all researchers working on the NPD data will only access the data via the SRS secure online system. We will apply for safe room connectivity to have SRS remote connectivity access.
- Safe outputs: All outputs will be checked by the ONS team to ensure that the outputs do not allow identification of individuals. Outputs will be checked against the Intended Permitted Outputs and be subject to standard ONS disclosure rules.
- Safe data: the data request includes data variables of identifiability risk level 3 (PMR), as the DfE will match the data we collect with the NPD data. The PMR (meaningless identifier) replaces the UPN and pupil name when the data are matched and then archived to minimise the risks of identification. Our researchers will only analyse de-identified data in the SRS.

Rights and Retention Periods

Parents can withdraw their child from the evaluation and/or from their data being processed, until it is added to the EEF archive. Should they withdraw from the programme or evaluation, NFER will still use the evaluation data that the school has provided up to that point and link it to NPD unless the parent indicates otherwise. If at any time, parents wish to withdraw their data or have errors corrected in it, contact details are provided in the Privacy Notices for who to contact about this.

As noted in the grant agreement, three months after the publication of the evaluation report, all of the pseudonymised matched data will be added to the EEF archive, which is managed by FFT on behalf of EEF and hosted by the ONS. This will enable the EEF and other research teams to use the pseudonymised data as part of subsequent research through the ONS Approved Researcher Scheme, including analysing long-term outcomes through the National

Pupil Database. This data may also be linked to other research datasets for the purpose of educational research.

NFER will securely delete any personal data relating to the evaluation one year after the publication of the final report, currently expected to be July 2023.

Data Controller and Processing Roles

The Department for Education (DfE) is the data controller and makes decisions about how personal data is used in the evaluation. EEF and OxEd and Assessment. Ltd are the data processors and NFER is the data sub-processor.

Personnel

Name	Affiliation	Roles and Responsibilities
Jack Worth	NFER	Project Director – responsible for overall quality of project delivery
Aarti Sahasranaman	NFER	Project Leader – responsible for day-to-day management of the project
Andrew Smith	NFER	Design Lead – responsible for design of impact evaluation and QA of analysis
Kathryn Hurd	NFER	Operations Lead – responsible for recruitment of NELI schools for impact evaluation, contacting schools to coordinate data collection, delivering incentive strategy
Jishi Jose	NFER	Project Manager – responsible for overseeing day-to-day running of operations
Ruth Staunton	NFER	Study statistician – responsible for preliminary and main analyses
Charles Hulme	OxEd and Assessment	CEO of OxEd and intervention developer. Year 2 NELI roll out delivery
Gillian West	OxEd and Assessment	Year 2 NELI roll out delivery. Director of the Year 2 NELI evaluation at OxEd and responsible for data extraction strategy
Joe Lowe	OxEd and Assessment	Year 2 NELI roll out delivery. Evaluation data extraction and data sharing
Sarah Hearne	OxEd and Assessment	Year 2 NELI roll out delivery. Project Manager – responsible for day-to-day management of Year 2 NELI evaluation at OxEd

Risks

Risk	Assessment	Controls, countermeasures and contingencies
Schools do not complete endpoint LanguageScreen assessments for all pupils who have a baseline assessment, instead focusing only on NELI pupils	Likelihood: High **Impact**: High	NFER and OxEd have created a coordinated communications strategy to send targeted reminders to schools to assess all reception pupils. Schools have also been notified that assessing the vast majority of reception pupils is a requirement for them to be eligible to earn up to £250 in incentive payments.
Schools do not indicate NELI status of pupils, thus affecting impact analysis	**Likelihood**: Moderate **Impact**: High	OxEd share weekly school-level data with NFER indicating the number of reception pupils marked as receiving NELI. NFER use these data to send tailored reminders to schools to complete and upload the data. Completion of the NELI indicator is also a requirement for schools to be eligible for the incentive payment.
Sample size not achieved	**Likelihood**: Moderate **Impact**: Moderate	547 schools have been recruited to the impact evaluation (with pupil-level data at August 2022 for 406 schools), higher than the required sample size of 468 schools. However, risks 1 and 2 above could reduce the sample of schools eligible to be included in the evaluation. Mitigations for the risks have been mentioned above. As part of the preliminary analysis, we are also undertaking a revised estimation of the sample size based on factors including the degree of fuzziness observed in the data. Furthermore, where schools have completed endpoint LanguageScreen assessments but not the NELI indicator, additional data collection is currently ongoing for the NELI indicator data.
Delays to project timelines because of additional data collection in autumn term of 2022	**Likelihood**: Low/Moderate **Impact**: Moderate	According to the original project timelines, data collection in schools was to be completed by July, 2022, following which OxEd would begin the process of extracting the final dataset to be shared with DfE by October, 2022. Since additional data collection is yet to be completed in schools, there could be some slippage of timelines. Although all the data has not yet been collected, OxEd have been refining their data extraction process in the meantime, with the result that any delays to sharing data with DfE may be minimised.

Timeline

Date	Activity	Staff responsible/ leading
March - May 2022	• Project set-up, development of recruitment materials, grant agreement sign off	NFER leading, but all staff involved
June 2022	• Receive list of schools who want to be contacted for impact evaluation from OxEd • Approach schools to bring them onboard • Remind participating schools to complete endpoint LanguageScreen assessments of all reception pupils and NELI status indicator • Receive weekly school-level data update from OxEd and monitor completion of assessments and NELI indicator data	NFER OxEd to provide school list and weekly school-level data
July 2022	• Monitor school-level data to assess completion of assessments and NELI indicator data • Send tailored communications to schools reminding them to complete missing assessments/NELI indicator data • NFER and OxEd agree on data spec for pupil-level data for preliminary analysis	NFER OxEd to provide weekly school-level data
August – September 2022	• Receive pupil-level data for preliminary analysis to assess viability of RD and to identify schools eligible to receive incentives • QA of pupil-level data • Additional data collection for NELI indicator data in schools • NFER and OxEd agree on data spec for pupil-level data for main analysis • Submit NPD data request	NFER OxEd to extract and share pupil-level LanguageScreen data
October 2022	• Submit draft of study plan • OxEd to share final de-identified pupil-level dataset with NFER • OxEd to submit final pupil-level dataset to DfE for matching of NPD variables • NFER to submit school-level dosage metric to DfE to combine with pupil-level data	NFER, OxEd
November - December 2022	• Finalise and publish study plan	NFER
December 2022 – February 2023	• Data linking with NPD • Complete preliminary analysis • Main analysis • Start report writing	NFER
March 2023	• Submit first draft of report	NFER
April – July 2023	• Report revisions and publication	NFER

References

Bowyer-Crane, C., Bonetti, S., Compton, S., Nielsen, D., D'Apice, K., and Tracey, L. (2021). The impact of Covid-19 on School Starters: Interim briefing 1. Available at: https://educationendowmentfoundation.org.uk/public/files/Impact_of_Covid19_on_School_Starters_-_Interim_Briefing_1_-_April_2021_-_Final.pdf (Accessed: 7 October 2022).

Bowyer-Crane, C., Snowling, M.J., Duff, F.J., Fieldsend, E., Carroll, J.M., Miles, J., Götz, K. and Hulme, C. (2008) 'Improving early language and literacy skills: Differential effects of an oral language versus a phonology with reading intervention', *Journal of Child Psychology and Psychiatry*, 49(4), pp.422-432. Available at: https://doi.org/10.1111/j.1469-7610.2007.01849.x

Cattaneo, M.D., Idrobo, N. and Titiunik, R. (2019) *A practical introduction to regression discontinuity designs: Foundations*. Cambridge University Press. Available at: https://www.cambridge.org/core/elements/practical-introduction-to-regression-discontinuity-designs/F04907129D5C1B823E3DB19C31CAB905 (Accessed: 7 October 2022).

Cattaneo, M.D., Idrobo, N. and Titiunik, R. (2020) *A practical introduction to regression discontinuity designs: Extensions*. Cambridge University Press. Available at: https://cattaneo.princeton.edu/books/Cattaneo-Idrobo-Titiunik_2018_CUP-Vol2.pdf (Accessed: 7 October 2022).

Cattaneo, M.D., Jansson, M. and Ma, X. (2018) 'Manipulation testing based on density discontinuity', *The Stata Journal*, 18(1), pp. 234–261. Available at: https://doi.org/10.1177/1536867X1801800115.

Cattaneo, M.D., Keele, L. and Titiunik, R. (2021) *Covariate adjustment in regression discontinuity designs*. Available at: http://arxiv.org/abs/2110.08410 (Accessed: 7 October 2022).

Cattaneo, M.D., Keele, L., Titiunik, R. and Vazquez-Bare, G. (2016) 'Interpreting regression discontinuity designs with multiple cutoffs', *Journal of Politics*, 78(4), pp. 1229–1248. Available at: https://doi.org/10.1086/686802.

Cattaneo, M.D. and Titiunik, R. (2022) *Regression discontinuity designs*, p. 34. Available at: https://titiunik.mycpanel.princeton.edu/papers/CattaneoTitiunik2022-ARE.pdf (Accessed: 7 October 2022).

Cattaneo, M.D., Titiunik, R. and Vazquez-Bare, G. (2019) 'Power calculations for regression-discontinuity designs', *The Stata Journal: Promoting communications on statistics and Stata*, 19(1), pp. 210–245. Available at: https://doi.org/10.1177/1536867X19830919.

De la Cuesta, B. and Imai, K., 2016. Misunderstandings about the regression discontinuity design in the study of close elections. Annual Review of Political Science, 19(1), pp.375-396.

Deke, J. and Dragoset, L. (2012) Statistical power for regression discontinuity designs in education: Empirical estimates of design effects relative to Randomized controlled trials, p. 24. Available at: https://files.eric.ed.gov/fulltext/ED533141.pdf (Accessed: 7 October 2022).

Department for Education and Ford, V. (2021) *Every school with reception class offered early language training*, *GOV.UK*. Available at: https://www.gov.uk/government/news/every-school-with-reception-class-offered-early-language-training (Accessed: 7 October 2022).

Dimova, S., Culora, A., Disley, E., and Gilder, L. (2021). Study Plan: The Nuffield Early Years Language Intervention Scale-Up Evaluation. Available at: https://d2tic4wvo1iusb.cloudfront.net/documents/pages/projects/NELI-Scale-Up-Study-Plan-Nov-2021-Final.pdf?v=1637662159 (Accessed: 7 October 2022).

Dimova, S., Illie, S., Brown, E., Rosa, Broeks, M., Culora, A. and Sutherland, A. (2020) *EEF-Evaluation-report-Nuffield-early-language-intervention.pdf*. Available at: https://www.nuffieldfoundation.org/wp-content/uploads/2016/02/EEF-Evaluation-report-Nuffield-Early-Language-Intervention.pdf (Accessed: 7 October 2022).

Disley, E., Nightingale, M., Haider, A. and Amodeo, G. (Forthcoming) *The Nuffield Early Language Intervention Scale-Up Evaluation Report – Year 2.*

Duff, F.J., Reen, G., Plunkett, K. and Nation, K. (2015) 'Do infant vocabulary skills predict school-age language and literacy outcomes?', J Child Psychol Psychiatry, 56, pp. 848–856. Available at: https://www.psy.ox.ac.uk/publications/502935 (Accessed: 7 October 2022).

EEF (2018). Statistical analysis guidance for EEF evaluations. Available at: https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Protocol_or_SAP/EEF_statistical_analysis_guidance_2018.pdf (Accessed: 7 October 2022).

Feinstein, L. and Duckworth, K. (2006) *Development in the early years: its importance for school performance and adult outcomes*. Available at: https://discovery.ucl.ac.uk/id/eprint/10005970/1/Feinstein2006Development.pdf (Accessed: 7 October 2022).

Fernald, A., Marchman, V.A. and Weisleder, A. (2013) 'SES differences in language processing skill and vocabulary are evident at 18 months', *Developmental Science*, 16(2), pp. 234–248. Available at: https://doi.org/10.1111/desc.12019.

Francis, B. (2022) 'EEF blog: Understanding the impact of COVID on learning', *Understanding the impact of COVID on learning*, 18 May. Available at: https://educationendowmentfoundation.org.uk/news/eef-blog-understanding-the-impact-of-covid-on-learning (Accessed: 7 October 2022).

Fricke, S., Bowyer-Crane, C., Haley, A.J., Hulme, C. and Snowling, M.J., (2013) 'Efficacy of language intervention in the early years', *Journal of Child Psychology and Psychiatry*, 54(3), pp.280-290. Available at: https://doi.org/10.1111/jcpp.12010

Fricke, S., Burgoyne, K., Bowyer-Crane, C., Kyriacou, M., Zosimidou, A., Maxwell, L., Lervåg, A., Snowling, M.J. and Hulme, C. (2017) 'The efficacy of early language intervention in mainstream school settings: A randomized controlled trial', *Journal of Child Psychology and Psychiatry*, 58(10), pp.1141-1151. Available at: https://doi.org/10.1111/jcpp.12737

Hedges, L.V. (1981) 'Distribution theory for Glass's estimator of effect size and related estimators', *Journal of Educational Statistics*, 6(2), pp. 107–128. Available at: https://doi.org/10.2307/1164588.

Heiss, A. (2020) *Fuzzy regression discontinuity*, *Program Evaluation*. Available at: https://evalf20.classes.andrewheiss.com/example/rdd-fuzzy/ (Accessed: 7 October 2022).

Home (no date). Available at: http://nfernet.nfer.ac.uk/ (Accessed: 7 October 2022).

Hulme, C. and Snowling, M.J. (2015) 'Learning to read: what we know and what we need to understand better', *Child development perspectives*, 7(1), pp. 1–5. Available at: https://doi.org/10.1111/cdep.12005.

Law, J., Todd, L., Clark, J., Mroz, M. and Carr, J. (no date) *Early language delays in the UK*. Available at: https://resourcecentre.savethechildren.net/document/early-language-delays-uk/ (Accessed: 7 October 2022).

McKensie, D. (2016) 'Power calculations for regression discontinuity evaluations: Part 1', 6 September. Available at: https://blogs.worldbank.org/impactevaluations/power-calculations-regression-discontinuity-evaluations-part-1 (Accessed: 7 October 2022).

Powell, L. (2021) *The Impact of COVID-19 for pupils using English as an additional language*. Available at: https://blog.insidegovernment.co.uk/schools/covid-19-impact-for-eal-pupils (Accessed: 7 October 2022).

Roulstone, S., Law, J., Rush, R., Clegg, J. and Peters, T. (2011) 'Investigating the Role of language in children's early educational outcomes'. American Psychological Association. Available at: https://doi.org/10.1037/e603032011-001.

U.S. Department of Education (2022) *What works clearinghouse procedures and standards handbook, version 5.0*, p. 234. Available at: https://ies.ed.gov/ncee/wwc/Docs/referenceresources/Final_WWC-HandbookVer5.0-0-508.pdf (Accessed: 7 October 2022).

West, G., Snowling, M.J., Lervåg, A., Buchanan-Worster, E., Duta, M., Hall, A., McLachlan, H. and 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, 62(12), pp.1425-1434.

West, G., Lervåg, A., Snowling, M.J., Buchanan-Worster, E., Duta, M. and Hulme, C. (2022) 'Early language intervention improves behavioral adjustment in school: Evidence from a cluster randomized trial', *Journal of School Psychology*, 92, pp. 334–345. Available at: https://doi.org/10.1016/j.jsp.2022.04.006.

Whitehurst, G.J. and Lonigan, C.J. (1998) 'Child development and emergent literacy', *Child Development*, 69(3), pp. 848–872. Available at: https://doi.org/10.2307/1132208.

Appendix A: Initial preliminary analysis

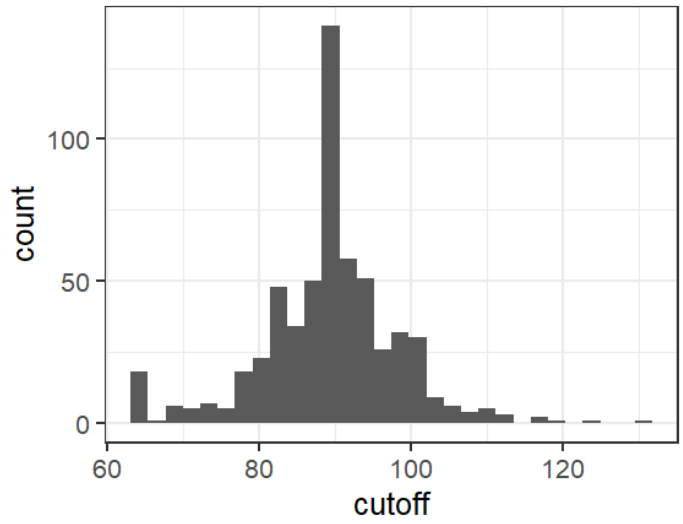


Figure 5: Histogram of preliminary analysis pre intervention LanguageScreen score cutoffs

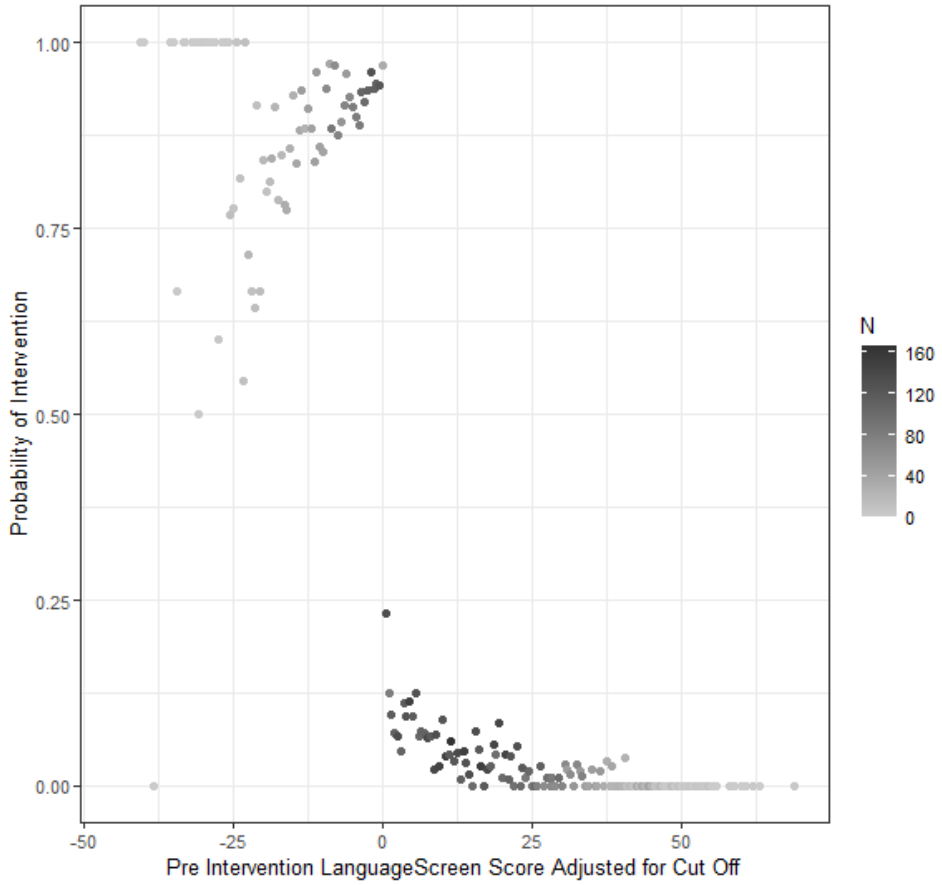


Figure 6: Probability of receiving intervention[10]

[10] Using cutoff adjusted pre-intervention LanguageScreen score (initial preliminary data analysis)

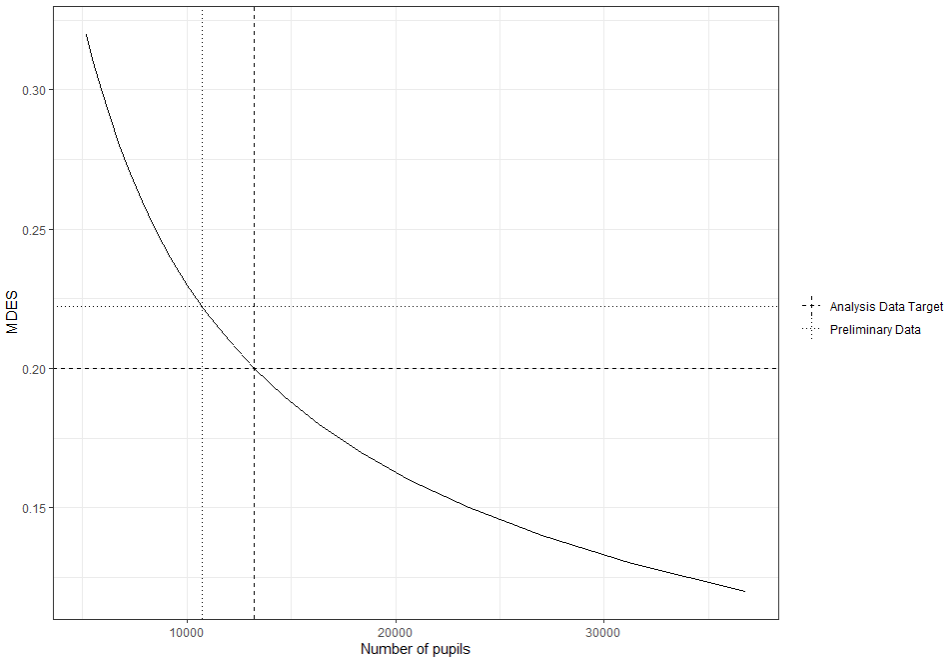


Figure 7: Minimum detectable effect size against sample size. Sample size calculation using preliminary data.

Appendix B: Initial sample size calculations for school recruitment (without data)

	Pupils				Schools			
	MDES				MDES			
	0.200	0.230	0.280	0.385	0.200	0.230	0.280	0.385
RCT	555	420	285	152	14	11	7	4
SRD, RDDE=9	4995	3780	2565	1368	125	95	64	34
SRD, RDDE=14	7770	5880	3990	2128	195	147	100	53
SRD, RDDE=17	9435	7140	4845	2584	236	179	121	65
FRD, 0.9, RDDE=9	6144	4649	3155	1683	154	117	79	42
FRD, 0.9, RDDE=14	9557	7232	4908	2617	240	181	123	66
FRD, 0.9, RDDE=17	11605	8782	5959	3178	291	220	149	80
FRD, 0.7, RDDE=9	9890	7484	5079	2709	248	188	127	68
FRD, 0.7, RDDE=14	15385	11642	7900	4213	386	292	198	106
FRD, 0.7, RDDE=17	18681	14137	9593	5116	468	354	240	128
FRD, 0.5, RDDE=9	18681	14137	9593	5116	468	354	240	128
FRD, 0.5, RDDE=14	29060	21991	14923	7959	728	551	374	199
FRD, 0.5, RDDE=17	35287	26704	18120	9664	884	669	454	242

Appendix C: Final pupil-level data (for analysis)

LanguageScreen data (from OxEd)

Variable name	Description
school_nfer	NFER school number
school_urn	URN
school_name	School Name
school_postcode	School Postcode
pupil_upn	UPN
pupil_meaningless_identifier	Pupil Meaningless Identifier
pupil_firstname	Pupil First Name
pupil_lastname	Pupil Last Name
class_name	School Class
pupil_date_of_birth	Date of Birth
pupil_gender	Gender
pupil_eal	EAL Status
intervention_received	Recieved NELI programme (yes/no)
approx_neli_completion_date	Completion date of NELI programme
pre_neli_assessment_date	Baseline Assessment Date
pre_neli_standard_score_ev	Baseline Expressive Vocabulary Score
pre_neli_standard_score_rv	Baseline Receptive Vocabulary Score
pre_neli_standard_score_lc	Baseline Language Comprehension Score
pre_neli_standard_score_sr	Baseline Sentence Repetition Score
pre_neli_standard_score_ls	Baseline Language Screen Standard Score
pre_neli_percentile_score_ls	Baseline Percentile Score
pre_neli_assessment_id	Baseline Assessment ID
pre_neli_assessment_age_months	Assessment age at baseline
post_neli_assessment_date	Endline Assessment Date
post_neli_standard_score_ev	Endline Expressive Vocabulary Score
post_neli_standard_score_rv	Endline Receptive Vocabulary Score
post_neli_standard_score_lc	Endline Language Comprehension Score
post_neli_standard_score_sr	Endline Sentence Repetition Score
post_neli_standard_score_ls	Endline Language Screen Standard Score
post_neli_percentile_score_ls	Endline Percentile Score
post_neli_assessment_id	Endline Assessment ID
post_neli_assessment_age_months	Assessment age at endline
time_between_assessments_days	Time between assessment in days
time_between_assessments_weeks	Time between assessment in weeks

NPD data

1. PupilMatchingRefAnonymous_[term][yy]
2. Gender_[term][yy]
3. YearOfBirth_[term][yy]
4. MonthOfBirth_[term][yy]
5. EthnicGroupMajor_[term][yy]
6. EVERFSM_6_P_[term][yy]
7. EVERFSM_ALL_[term][yy]
8. FSMeligible_[term][yy]
9. LanguageGroupMajor_[term][yy]
10. PartTime_[term][yy]
11. SENprovisionMajor_[term][yy]
12. PrimarySENtype_[term][yy

School-level characteristics (derived from DfE's publicly available Get Information about Schools dataset)

School-level dosage and fidelity data (derived from Nuffield Foundation Ltd's TeachNELI delivery surveys and OxEd's Futurelearn training platform.