THE NUFFIELD EARLY LANGUAGE INTERVENTION

Addendum Report

March 2023

Authors: Merrilyn Groom, Elena Rosa Brown, Lydia Lymperis

The Education Endowment Foundation is an independent charity dedicated to breaking the link between family income and education achievement. We support schools, nurseries and colleges to improve teaching and learning for 2 – 19-year-olds through better use of evidence.

We do this by:

- **Summarising evidence.** Reviewing the best available evidence on teaching and learning and presenting in an accessible way.
- **Finding new evidence.** Funding independent evaluations of programmes and approaches that aim to raise the attainment of children and young people from socio-economically disadvantaged backgrounds.
- **Putting evidence to use.** Supporting education practitioners, as well as policymakers and other organisations, to use evidence in ways that improve teaching and learning.

We were set-up in 2011 by the Sutton Trust partnership with Impetus with a founding £125m grant from the Department for Education. In 2022, we were re-endowed with an additional £137m, allowing us to continue our work until at least 2032.

For more information about the EEF or this report please contact:

Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
SW1P 4QP

info@eefoundation.org.uk

www.educationendowmentfoundation.org.uk

Contents

About the Evaluator

The project was independently evaluated by a team from RAND Europe: Merrilyn Groom, Elena Rosa Brown, Lydia Lymperis.

Contact details:

Name: Elena Rosa Brown

Address: RAND Europe, Westbrook Centre, Milton Road, Cambridge, CB4 1YG

Email: Elena_Rosa_Brown@randeurope.org

Acknowledgements to Eliane Dufresne, Sasha Dimova, and Sonia Ilie who supported the research while at RAND. Also thank you to Camilla Nevill, Eleanor Stringer, and Sarah Tillotson at the EEF for supporting us, and to the University of Oxford team for working with us during the evaluation.

Executive summary

The project

Nuffield Early Language Intervention (NELI) is an early language programme for Reception pupils (aged 4–5), involving scripted individual and small-group language teaching sessions delivered by trained teaching assistants (TAs) for 20-weeks.

Studies of NELI have demonstrated positive effects on children's language skills. A previous trial funded by the EEF in 34 schools found that pupils involved in a 30-week version of NELI while in Nursery and Reception made an additional 4 months' progress in language skills, and that gains in language skills were maintained six months later. The results of the effectiveness trial, involving 1,156 pupils in 193 schools, found that children receiving the NELI programme made the equivalent of three additional months' progress in oral language skills, on average, compared to children who did not receive NELI. This result had a very high security rating of five out of five padlocks.

This evaluation, an addendum report to the effectiveness trial, focused on exploring the impact of the NELI language intervention (approximately two years later) on reading, including early word reading, reading fluency and reading comprehension. It also examined the extent to which language impact was sustained using a latent measure of language.

A number of tests were administered to pupils for the purposes of the longitudinal follow-up, including the Renfrew Action Picture Test (RAPT) and LanguageScreen, to assess language, and three York Assessment for Reading Comprehension (YARC) subtests (early word reading, reading comprehension, reading fluency). Children were between the ages of 6 and 7 at follow-up.

The intension had been for external testers to assess the pupils when they were 5 or 6 years old at the end of year 1, one year after delivery of the NELI intervention was complete. However, the Covid-19 pandemic and partial closure of schools led to a decision to delay the follow up assessments to July 2021 with some assessment complete in Autumn 2021. The pandemic's continued disruption to school life resulted in a high rate of attrition (55.5%), meaning a smaller sample at follow-up, and the resulting relatively high MDES. The results of this report therefore should be interpreted with caution.

Key conclusions
Much of the gains to oral language as a result of NELI persist over time for those pupils included in the follow-up analysis.
The effect of NELI on early word reading persists over time for those pupils included in the follow-up analysis.
NELI appears to lead to improved reading comprehension for those pupils included in the follow-up analysis. This is in line with the logic model with improvements in language leading to improvements in reading comprehension.
There appears to be a larger effect of NELI on language for FSM-eligible pupils compared to non-FSM-eligible pupils. However, this finding should be interpreted with caution, given the low statistical power.

Additional findings

Exploratory analysis suggests that two school years later, NELI has an impact on reading comprehension and early word reading. In contrast, NELI has less of an impact on reading fluency. Both of these findings are in line with the hypothesised outcomes in the logic model (Figure 1).

Exploratory analysis also found that NELI continues to have a sustained influence on oral language in children included in the follow-up, aged between 6 and 7. This conclusion is further supported by the sensitivity analysis, which shows that much of the impact seen in the effectiveness trial persists over time.

The FSM subgroup analysis, which looked at the extent to which NELI had a differential impact for FSM-eligible pupils, suggests that there is an observable positive difference in oral language between treatment and control, and this effect is larger for FSM pupils compared to non-FSM pupils. However, this analysis was notably underpowered. The effectiveness trial did not conduct an FSM subgroup analysis on oral language as the data was not available, with follow-up presenting the first opportunity to do so.

The English as an additional language (EAL) subgroup analysis, which was also notably underpowered, indicates that NELI effects on EAL pupils' language skill development for those pupils that were analysed at follow-up persisted over time. This is in line with the effectiveness trial, which also showed a positive effect on EAL learners.

Results should be interpreted with caution given the very high attrition across the sample and the change in some of the outcome measures. As such, these analyses were intended to be exploratory.

Table 1: Summary of findings from analysis

Outcome/ group	**Effect size (95% confidence Interval)**	**Estimated months' progress**	**No. of pupils**	***p*-value**	**EEF cost rating**
Early word reading	0.18 (0.02, 0.39)	2	514	N/A	N/A
Reading comprehension	0.17 (0.00, 0.35)	2	513	N/A	N/A
Reading fluency	0.04 (−0.14, 0.20)	0	513	N/A	N/A
Early language	0.18 (0.02, 0.35)	2	507	N/A	N/A

Introduction

Intervention

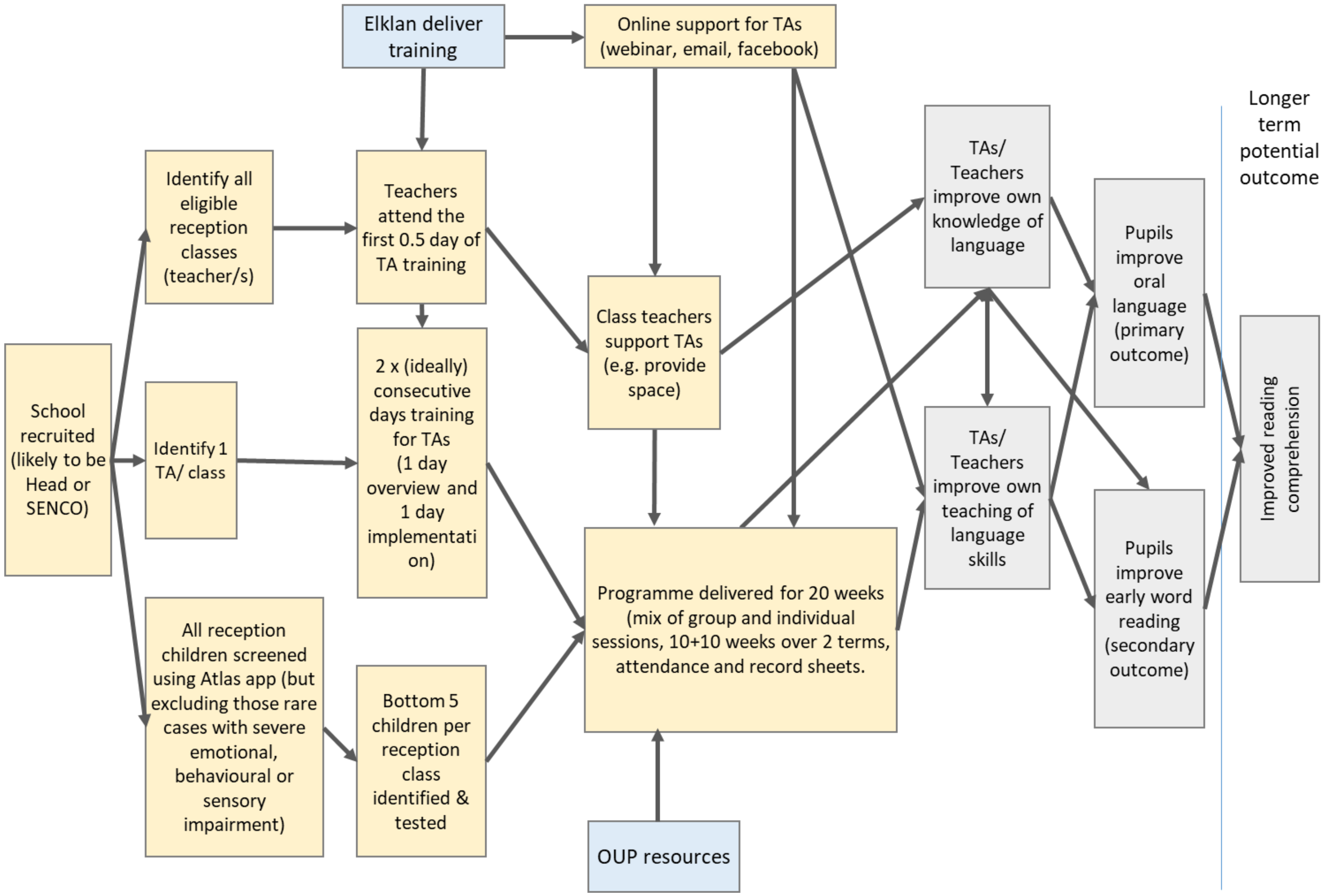
The original report, The Nuffield Early Language Intervention: Evaluation Report (Dimova et al., 2020), provides a full description of the design, implementation and findings of the effectiveness trial on intensive language support for young children with poor spoken language skills.

The Nuffield Early Language Intervention (NELI) programme is designed to improve the language skills of Reception pupils (aged 4–5) and involves scripted individual and small-group language teaching sessions delivered by trained teaching assistants (TAs). NELI is published by Oxford University Press (OUP) and was delivered in this trial by the University of Oxford in partnership with Elklan. As can be seen in Figure 1, the 20-week intervention consists of three 30-minute small group sessions and two 15-minute individual sessions each week. The children selected to participate in this trial were the five children in each classroom who obtained the lowest scores on a school-administered app-based assessment of oral language skills (LanguageScreen). The programme focuses on developing children's narrative, vocabulary and listening skills, while phonological awareness and letter–sound knowledge are also reinforced during the second half of the programme. Training includes an initial 2-day course for TAs, followed by a half-day workshop half-way through the programme in addition to ongoing online support. Teachers also receive an initial half-day of training so that they can understand the programme and support TAs.

The intervention was evaluated using a randomised controlled trial; 193 schools participated, 97 of which received the programme while 96 did not, the latter forming the control group. The process evaluation involved observations of the training, surveys with school staff, case studies in 6 schools (which used interviews with school staff), and analyses of TA training logs and online resource use. This project was co-funded with the financial support of Intermediate Capital Group (ICG). The trial took place between June 2018 and November 2019. Follow-up testing took place between May and October 2021, with testing falling either side of the summer holiday due to the Covid-19 pandemic.

For an in-depth description of the intervention and evaluation please refer to the original report (Dimova et al., 2020).

Figure 1: NELI logic model


Elklan deliver training
Online support for TAs (webinar, email, facebook)
Identify all eligible reception classes (teacher/s)
Teachers attend the first 0.5 day of TA training
School recruited (likely to be Head or SENCO)
Identify 1 TA/ class
2 x (ideally) consecutive days training for TAs (1 day overview and 1 day implementati on)
Class teachers support TAs (e.g. provide space)
All reception children screened using Atlas app (but excluding those rare cases with severe emotional, behavioural or sensory impairment)
Bottom 5 children per reception class identified & tested
Programme delivered for 20 weeks (mix of group and individual sessions, 10+10 weeks over 2 terms, attendance and record sheets.
OUP resources
TAs/ Teachers improve own knowledge of language
TAs/ Teachers improve own teaching of language skills
Pupils improve oral language (primary outcome)
Pupils improve early word reading (secondary outcome)
Longer term potential outcome
Improved reading comprehension

Evaluation objectives

This follow-up study has three main research questions it wished to explore:

1. To what extent does the impact seen on language and early reading in the NELI effectiveness trial transfer to reading comprehension and fluency?
2. To what extent do language differences between pupils who took part in NELI and 'business-as-usual' pupils persist?
3. Are there differential effects of NELI on the oral language of FSM-eligible pupils compared to non-FSM-eligible pupils?

For the purposes of this report, we will refer to this as the follow-up evaluation, and the previous trial as the main trial.

As the original study was not designed for longitudinal follow-up, this follow-up evaluation should be considered exploratory and interpreted with caution. Due to the high attrition rate (55.5%), analysis is both underpowered and at greater risk of bias. As the analysis is exploratory, this report does not place much weight on measures of statistical significance and, accordingly, does not report MDES in line with EEF guidance on longitudinal analysis.[1] Additionally, Covid-19 presented problems for conducting follow-up testing, which resulted in a change to the outcome measures between initial endline and follow-up. This means the primary outcome used at endline could not be replicated at follow-up. As the report was exploratory in nature, the evaluation team did not publish a statistical analysis plan (SAP). Code and output has been included in the appendix B, C and D, instead, to allow for scrutiny and replication.

Project team

The intervention was developed and implemented by the University of Oxford. Elklan was a delivery partner, with Elklan trainers being trained by the delivery team and then training school staff. The overall responsibility for training school staff and providing support to all schools as per the terms of the sign-up information sheet ultimately rested with the University of Oxford team. The University of Oxford team was also responsible for organising and administering baseline testing. In addition, they provided the key indicators of implementation and compliance with the intervention. The delivery team at Oxford also provided input into the design of the trial. The delivery team at Oxford was led by Professor Charles Hulme and Dr Gillian West, and also involved Professor Maggie Snowling, Ms Denise Cripps, and Dr Elizabeth Worster. The follow-up testing was run by the team at the University of Oxford led by Professor Charles Hulme and Dr Gillian West, and included Mariela Rios-Diaz and Caroline Korell, with assistance from Elklan, who managed the re-recruitment of schools from the main trial.

The current evaluation team in RAND Europe includes Elena Rosa Brown (current project leader), Merrilyn Groom (current project manager, co-lead analysis) and Lydia Lymperis (co-lead analysis). They have been supported by a team of current and former staff at RAND Europe, including: Eliane Dufresne (formerly project management and analysis), Dr Sonia Ilie (project leader, formerly RAND Europe), Dr Sashka Dimova (project lead, project management, field work and analysis, formerly RAND Europe), Miriam Broeks (field work and analysis, formerly RAND Europe), Dr Andreas Culora (project management and analysis, formerly RAND Europe), Giulia Lanfredi (report writing, formerly RAND Europe), Eleftheria Iakovidou (field work, formerly RAND Europe), Dr Susie Lee (field work, formerly RAND Europe), Natalie Picken (analysis), and Dr Alex Sutherland (project leader, formerly RAND Europe). The evaluation team also benefitted from advice and quality assurance from Dr Monica Melby-Lervåg and Dr Arne Lervåg (University of Oslo).

The evaluation was conducted independently by RAND Europe. The evaluator was responsible for the outcome and process evaluation, trial design (with input from the Oxford team based on prior NELI evaluation experience and, in response to a prior NELI trial protocol by Dunne and Miller, 2017), analysis, reporting and quality assurance of the evaluation. The Oxford team initially pre-registered the trial independently of the evaluator; this registration was then amended by the independent evaluator to reflect the final design.

[1] longitudinal_guidance.pdf (educationendowmentfoundation.org.uk)

Methods

Trial design

The trial was planned and executed as a stratified, two-arm, cluster-randomised controlled trial (cRCT) across 193 schools and a total of 240 Reception classrooms recruited from thirteen geographical regions: Bristol, Cornwall, Durham, Essex, Hertfordshire, London, Manchester, North Tyneside, Northamptonshire, Surrey, Warwickshire and surrounds, Wolverhampton, and the North West of England. The areas where recruitment took place were selected based on available contacts, and to ensure an equal spread across urban and rural regions.

Schools were assigned to either treatment (NELI) or control (business as usual). All schools signing up had a 50:50 chance of being assigned to the treatment group within each geographical cluster. Schools randomly assigned to treatment received training and support to implement NELI in the 2018–19 academic year. Schools randomly allocated to control continued with business as usual during the same academic year. As an incentive, control schools received financial compensation of £1,000: £500 post-randomisation and £500 upon completion of outcome testing.

Participant selection and sample size

Schools

Schools meeting the following criteria were eligible for inclusion in the study:

1. has not previously delivered NELI
2. has above-average free school meal (FSM) eligibility2
3. is willing to be randomly assigned to intervention or 'business as usual' at the level of the school
4. is willing to engage with the intervention and implement it with the pupils identified by screening
5. is willing to provide child background information to the evaluation team
6. is willing to facilitate baseline and post-intervention data collection.

Eligible schools were accepted into the trial once they had completed the required paperwork and prerequisite tasks, and after signing a Memorandum of Understanding (MoU). The MoU specified that once schools agreed to participate, the expectation was that final outcome testing of children would be allowed, even if the school were to withdraw from the intervention. All schools that fulfilled the inclusion criteria and provided the paperwork and prerequisite tasks were eligible for randomisation. 193 schools were randomised, with 97 randomly assigned to treatment and 96 randomly assigned to control.

Children

All children in Reception classrooms within eligible schools, approximately 6,000 children overall, were included in the screening process. Teachers had the option to exempt children with such severe hearing, vision or behavioural issues that the child would not be able to access the programme. Children in the participating Reception classes took part in the screening, during which their oral language skills were measured using LanguageScreen. LanguageScreen was developed by the delivery team during the NELI efficacy trial, and is used to select the children who should receive NELI. It is a language test (and accompanying digital application for the administration of the test) that assesses four core language skills: expressive vocabulary, receptive vocabulary, sentence repetition and listening comprehension.

Based on the score derived from the use of LanguageScreen, the five children in each participating Reception classroom with the lowest composite language scores were selected for inclusion in the intervention and further testing. Further individual testing of the children selected as eligible for NELI was completed by Elklan prior to randomisation, by trained testers who were speech and language therapists (see further details on tests in the

[2] Around 13.7% pupils were eligible for FSM in 2018.

Outcome measures section below). Randomisation at the school level resulted in 585 pupils randomly allocated to treatment and 571 pupils randomly allocated to control (1,156 pupils in total).

Outcome measures

Original outcome measures in the effectiveness trial

The following tests were administered in the effectiveness trial at baseline and endline by Elklan testers who were blind to allocation:

1. Clinical Evaluation of Language Fundamentals (CELF) Preschool 2UK: Recalling sentences
2. CELF Preschool 2UK: Expressive Vocabulary
3. Renfrew Action Picture Test (RAPT) information
4. RAPT grammar.

The primary outcome measure in the effectiveness trial was constructed as an amalgamation of the four post-tests, using confirmatory factor analysis (CFA) as part of a structural equation modelling (SEM) approach.

Two secondary outcome measures were captured as part of the effectiveness trial:

1. York Assessment of Reading Comprehension (YARC) early word reading test (early reading skills)
2. LanguageScreen.

Follow-up outcome measures

The intention had been for external testers to administer the tests that were used in the effectiveness trial with the addition of the YARC reading comprehension and YARC reading fluency tests to measure reading. Testing was to be done in the same schools in July 2020, with the same pupils as part of this follow-up. However, owing to schools restricting visitors as a result of the Covid pandemic, the decision was made by EEF, the University of Oxford and RAND to ask school staff to administer the tests. As a result of this, the CELF was excluded, as the children would have been on the upper end of the age limit for the test (aged 6–11). Testing with the revised test battery was delayed to May 2021, and eventually completed in October 2021 (see Table 1 for a tally of pupils tested in each month). This means that the testing window spanned two academic years, with 93% of children assessed while they were in the last term of Year 2 and the remaining 7% of children assessed in the first term of Year 3, with a break over summer. In total 90% of schools participated in testing in the summer term of 2021, and 10% in the autumn of 2021. This is problematic for a number of reasons. The extended timeline increases the risk of pupils' 'age out' of the tests becoming older than is suitable. It also makes comparability difficult as there is potentially a difference of 4 months in age between pupils who sat the test in the beginning, compared to those that sat the test at the end. Finally, summer learning loss could have further exacerbated differences between pupils who were measured in July and those who were measured in September and October. However, given the constraints of outcome testing, extending testing was the most pragmatic option. To understand if age at time of test was related to outcomes, we undertook an extra sensitivity analysis (see *Analysis* subsection).

Schools were sent testing packs, including stimulus materials and paper record forms and a link to a Qualtrics data entry survey. A detailed 'school assessor' (i.e., teacher and/or TA) training course was developed by the University of Oxford team to aid schools with testing and was shared via the FutureLearn platform. School assessors could only administer the tests once they had successfully completed the training.

LanguageScreen was scored automatically in the App. For RAPT and YARC, school assessors were asked to record children's verbatim responses and upload these to the University of Oxford team via Qualtrics for scoring by the research team. To ensure accuracy of raw data, schools were also sent a hand-held audio recorder and asked to record all testing sessions. Schools were asked to listen to recorded sessions to ensure their Qualtrics transcription of raw test data were accurate. On completion of testing, schools were asked to return the audio recorders to the University of Oxford team, who then checked a random sample for transcription accuracy.

Table 1: Follow-up testing timeline

Follow-up testing timeline		
Month of testing	**Number of children tested and analysed**	**Number of schools**[1]
May 2021	55	14
June 2021	280	63
July 2021	142	46
September 2021	36	13
October 2021	1	1

[1] **Some schools ran testing over multiple months**

Table 2 highlights the differences between tests in the original dataset and the follow-up study. Given the exploratory nature of this follow-up, there are no primary or secondary outcomes.

Table 2: Outcome measures across original and follow-up studies

Outcome		Original effectiveness study	Follow-up study
Reading outcomes	Variable	Early word reading Alternative latent oral language	Early word reading Reading fluency Reading comprehension
	Measures (instrument, scale, source)	YARC early word reading test	YARC early word reading test YARC reading fluency YARC reading comprehension
Language outcome*	Variable	Latent oral language	Latent oral language
****Primary outcome in the effectiveness trial***	Measures (instrument, scale, source)	Primary latent language variable created using four language tests: **1** CELF: recalling sentences subtest **2** CELF: expressive vocabulary subset **3** RAPT: information **4** RAPT: grammar Secondary latent language variable created using the four LanguageScreen tests: **1** Language Screen: Listening Comprehension subtest **2** LanguageScreen: Expressive Vocabulary subtest **3** LanguageScreen: Receptive Vocabulary subtest **4** LanguageScreen: Sentence Repetition subtest	Latent language variable created using six language tests: **1** RAPT: information **2** RAPT: grammar **3** LanguageScreen: Listening Comprehension subtest **4** LanguageScreen: Expressive Vocabulary subtest **5** LanguageScreen: Receptive Vocabulary subtest **6** LanguageScreen: Sentence Repetition subtest

Reading outcomes

Given the evaluation's interest in exploring how NELI transferred into reading skills, additional subscales of the YARC were added at follow-up, specifically reading fluency and reading comprehension. Reading comprehension maps onto NELI's hypothesised long-term impact in the logic model (Figure 1). Reading fluency was included, despite not appearing in the logic model, as it is tested with the same passages as reading comprehension and is thus collected concurrently to reading comprehension. It was considered an appropriate potential outcome to explore so it was included in the analysis as an exploratory outcome of interest.

Language outcome

LanguageScreen was considered a suitable replacement for CELF by the delivery team, EEF and RAND. LanguageScreen, which is also designed to measure oral language, was measured at follow-up as well as at baseline and endline in the original effectiveness study. There were several advantages to using LanguageScreen – it measures both expressive vocabulary and sentence repetition, two aspects of oral language that were originally captured by CELF and, unlike, the CELF, is designed to be administered by teachers, which was a requirement due to schools restricting access to non-school staff as a result of Covid-19. It has been validated in a sample of 350,000 children (Hulme et al, pre-print). One potential criticism is that the measure was developed by the delivery team, which could introduce bias as the measure may be aligned to the intervention (i.e., by asking specific content questions covered by the NELI programme). However, the measure was designed to be a general screening tool, not specifically or solely for NELI.

As CELF was not administered at follow-up (see above), the structural equation model (SEM) had to be altered from that used in the effectiveness trial. The language outcome measure at follow-up is a latent factor variable, constructed from both RAPT variables and all of the LanguageScreen variables, using CFA that is broadly similar to the original SEM approach (see further discussion in the *Approach to SEM* section below). A stable SEM could not be created using the two RAPT variables alone, so to generate a latent oral language factor, the SEM had to also include the LanguageScreen variables. As LanguageScreen is a developer-led test, a number of robustness checks on the new oral language factor model were undertaken to ensure the results at follow-up are comparable with those found in the effectiveness report. As discussed below, the new latent model could replicate the original effect found in the effectiveness report to within three-hundredths, suggesting that the two models are comparable and that no additional bias was introduced by using LanguageScreen instead of CELF.

Analysis

The outcome analysis was undertaken on an intention-to-treat (ITT) basis. The analysis included all randomised schools in the groups to which they were randomly assigned initially, regardless of the treatment they actually received, withdrawal post-randomisation, or any deviations in programme implementation. This principle is essential to ensuring non-biased intervention effect estimation and compares outcome means for the treatment and comparison groups as they resulted from the randomisation procedure. Therefore, the ITT approach is inherently conservative as it captures the averaged effect of offering the intervention, regardless of whether or not the participants comply with assignment.

Analysis was not blind to allocation. Stata analysis code (and output) has been provided and data will be deposited in the archive to allow for potential future replication work.

Following the approach adopted in the initial effectiveness evaluation we explored the possibility of creating a latent language construct using SEM.

Approach to structural equation modelling (SEM)

Structural equation modelling (SEM) is a broad statistical technique that is routinely used to hypothesise, depict and then test relations among observed and latent variables, thereby making it possible to empirically test the measurement structure of specific theoretical constructs (Schumacker and Lomax, 2016). A latent variable is defined as a substantive theoretical construct that is only measured indirectly, with several indicators all capturing different aspects of that construct (Muthén, 2002). SEM encompasses statistical techniques such as CFA, regression analysis and path analysis. It allows for the estimation of a so-called factor score, which combines all underlying indicators making up the latent variable and their respective contributions to that variable.

In SEM and confirmatory factor analysis, model fit is assessed by means of a series of fit indices. The measurement model was assessed for fit (how well the observed variables combine to capture the underlying language construct) by using the recommendations of Hu and Bentler (1999) for cut-off criteria. In particular, the most often reported indices of fit are the comparative fit index (CFI), which compares the fit of a target model to the fit of a null (or independent) model, with good fit indicated by values above 0.95, and the root mean square error of approximation (RMSEA), which adjusts for model parsimony, with good fit indicated by values below 0.06.

Additionally, a chi-squared statistic, which assesses the overall fit and discrepancy between the sample and the fitted covariance matrices, is also usually reported. Non-significant values are indicative of good model fit in SEM, although established methodological literature suggests that, in large samples, the chi-squared test of exact fit will often reject the null hypothesis, even when the postulated model is only trivially false (Shi et al., 2019). Indeed, there is some disagreement in the field as to cut-off points and the indices to be reported; literature has explored in simulation studies how both fit indices are liable to lead to misinterpretation of model fit (Kenny, 2015) or can sometimes disagree with each other (Lai and Green, 2016). As such, we report the chi-squared statistic, but do not rely solely on chi-squared alone to assess the performance of the model given the concerns outlined here.

The main advantage of SEM over other analytical techniques that are commonly used to combine different subscales into a single dimension, such as computing a simple mean or summative score, is the fact that the factor score weights each subscale according to how salient it is to the concept being measured. Therefore, in cases where some indicators have much weaker loadings than others, a factor score is expected to be more reliable than a summative or mean score (Acock, 2013). All SEM analyses were carried out in Stata 17 using the **sem** command.

Analysis of reading outcome measures

We initially explored using SEM to estimate an underlying latent factor for reading from the three YARC subtests. However, SEM failed to achieve a good level of fit as measured by RMSEA, taking on a value of 0.099. In particular, reading fluency loaded negatively on the underlying latent factor, suggesting that improvements in reading fluency are negatively correlated with underlying reading ability. This may be linked to the presence of substantial floor effects, as is suggested by the histograms (see Appendix A); while YARC has been validated on this age group, it is possible that, since NELI was targeted at those pupils at the lowest end of the distribution at baseline, YARC reading fluency may not be sensitive enough to distinguish between pupils in the lower tail in this age group. The existence of floor effects alone, however, may not account for the poor fit, as YARC reading comprehension also exhibits some floor effects but still loads positively on the underlying latent factor. It could be that the SEM model is not capturing the hypothesised construct, so we proceed instead by analysing each of the variables in turn. Analysing reading comprehension and early word reading without the use of SEM is still in accordance with the logic model, as the logic model focuses on these two measures of reading, with no reference to reading fluency.

To estimate the impact of NELI on reading skills captured through YARC, we used a two-level multilevel model to account for clustering of pupils in schools on each of the three different subcomponents of YARC. This is in line with the EEF's stated preference for using multilevel modelling for clustered trial designs. The multilevel modelling framework can flexibly handle complex variation within/between schools (Hox, 1998; Snijders, 2005; Snijders and Bosker, 1994).

This reading outcome analysis consisted of model (1) below, for outcomes of pupils nested in schools, which matches the approach for the language outcome analysis.

$$Y_{ij} = \beta_0 + \mathrm{NELI}_j \tau_{YARC} + Z_j\beta_1 + X_{ij}\beta_2 + u_j + e_{ij} \tag{1}$$

where Y_{ij} is the YARC score (either early word reading, reading fluency or reading comprehension) for child i in school j; NELI_j is a binary indicator of the school assignment to intervention [1] or control [0]; Z_j are school-level characteristics, here the two stratifying variables of geographical location and number of classes per schools (as used for randomisation); X_{ij} represent characteristics at pupil level (pupil i in school j), specifically the baseline YARC assessment; u_j are school-level residuals ($u_j \sim i.i.d\ N(0, \sigma_u^2)$) and e_{ij} are individual-level residuals ($e_{ij} \sim i.i.d\ N(0, \sigma_e^2)$).

YARC reading fluency and reading comprehension were not measured at baseline, so we instead include YARC early word reading at baseline for all three secondary outcomes, as a proxy for baseline reading ability.

For the reading outcome analysis, τ_{YARC} is the average effect of the intervention on pupil outcomes in treatment schools compared to those in control schools when using the YARC measures for pre-test and post-test. This constitutes the reading outcomes of the trial.

Analysis of language outcome measure

SEM measurement models are flexible but must be underpinned by robust theoretical assumptions. As such, the starting point of the SEM analysis was the development of a conceptually driven measurement model that describes the underlying measurement structure of the oral language construct. The main theoretical assumption relevant for the development of this construct is that the six indicators (coming from the six tests above) capture aspects of a single underlying theoretical dimension – oral language skills – and so are likely to be highly correlated with each other. By modelling the underlying relationship between the tests and the latent construct of interest – oral language – we allow for the indicators to differ empirically in their relative loading onto the underlying latent factor. This approach accords with the approach taken in the effectiveness trial.

The resulting model for oral language has a good level of fit, according to indices of fit, with a RMSEA of 0.048, which is below the 0.06 threshold, and a CFI of 0.99, above the threshold of 0.95. Given this latent factor is also used as the pre-test outcome, the CFA model should ideally be a good fit at baseline as well. According to both measures, the baseline model also demonstrates an acceptable-to-good level of fit, with a RMSEA of 0.075 and a CFI of 0.983. Full details of the model are provided in Appendix C.

The procedure described above was applied to both the pre-test and post-test language construct. Once model fit was established, a factor score was generated for both the pre-test (baseline) and post-test (outcome) measure separately. Each of these two factor scores therefore captured the language latent variable in such a way that it could be included in the EEF-preferred multilevel model specification, as follows.

The language outcome analysis follows the same model used in the effectiveness evaluation, to ensure results are as comparable as possible. It is outlined in model (2) below, for outcomes of pupils nested in schools:

$$Y_{ij} = \beta_0 + \mathrm{NELI}_j\tau + Z_j\beta_1 + X_{ij}\beta_2 + u_j + e_{ij} \qquad (2)$$

where Y_{ij} is the language factor score for child i in school j; NELI_j is a binary indicator of the school assignment to intervention [1] or control [0]; Z_j is a vector of school-level characteristics, here the two stratifying variables of geographical location and number of classes per schools (as used for randomisation); X_{ij} represent characteristics at pupil level (pupil i in school j), specifically the baseline language factor score; u_j are school-level residuals ($u_j \sim i.i.d\, N(0, \sigma_u^2)$) and e_{ij} are individual-level residuals ($e_{ij} \sim i.i.d\, N(0, \sigma_e^2)$). Model (1) is known as a 'random intercepts' model because $\beta_{0\mathrm{j}} = \beta_0 + u_j$ is interpreted as the school-specific intercept for school j and $\beta_{0j} \sim i.i.d\, N(\beta_0, \sigma_u^2)$ is random (that is, it is assumed to be random).

Our target parameter, and the coefficient upon which the language outcome result of the trial draws, is τ, which is the average effect of the intervention on pupil outcomes in intervention schools compared to those in control schools, after accounting for the effect of the baseline assessment and the two stratification variables.

All analyses were performed in Stata, versions 15.1 onwards (Heß, 2017).

Missing data analysis

Missing data can arise from item non-response or attrition of participants at school and pupil levels. Attrition for this evaluation was substantial at follow-up, both due to school-level attrition (schools dropping out at follow-up) and pupil-level attrition (a higher proportion of pupils left the school at follow-up, alongside missingness due to usual illnesses and non-attendance).

We explored attrition across trial arms as a basic step to assess bias (Higgins et al., 2011). We provided cross-tabulations of the proportions of missing values on all baseline characteristics as well as on the primary outcome measures (see *Attrition* section below).

To understand the extent to which missingness was systematic, we modelled missingness at follow-up as a function of baseline covariates, including treatment. The analysis model for this approach mirrors the multilevel model given above (pupils clustered in schools), but the outcome is a binary variable identifying missingness (yes/no). The model used the following predictors: (i) at the pupil level, English as an additional language (EAL) status and latent oral language at baseline, and (ii) at the school level, school prior attainment at Key Stage 2 (KS2), proportion of pupils who have EAL or qualify for FSM, and Ofsted ratings.

Given missingness at follow up is over 55% but seemingly unrelated to observables, it seems likely that data is missing not at random (MNAR; see *Attrition* subsection below). We thus assumed data is MNAR, in which case neither multiple imputation (MI) or full-information maximum likelihood (FIML) would reduce the bias introduced by missingness. Even if we did assume data was missing at random (MAR), the scale of missingness at follow-up makes complete case the most robust approach (Jakobsen et al., 2017).

As a result, we revised our approach to missingness. We focused on describing missingness and examining patterns of missingness using a logit model to attempt to bound the primary outcome effect size (e.g., an upper or lower bound of the true effect size) based on any observed patterns of missingness, rather than in trying to use imputation and sensitivity analysis of the imputed model to generate alternative estimates of the effect on primary outcome.

Subgroup analyses

The study was not powered for meaningful subgroup analysis. However, as the effectiveness study found an impact of three months' progress on children who speak English as an additional language (EAL), a focus on the EAL subgroup is warranted. Therefore, as an exploratory modelling approach, EAL is incorporated into the analysis as a binary moderator variable, taking the value [1] if a child is EAL and [0] otherwise. The EAL indicator will then be interacted with treatment allocation to assess the conditional impact of NELI on EAL pupils. As this analysis is exploratory and underpowered, we do report point estimates and confidence intervals transformed into effect sizes but do not report significance tests/*p*-values.

The effectiveness trial did not conduct a subgroup analysis on pupils eligible for FSM, as the FSM data was not available for the age group in the original trial. This data is now available for those pupils who were present at follow-up, so a subgroup analysis exploring the effect of NELI on language outcomes at follow-up for FSM pupils was conducted. As per EEF guidance, FSM was explored both through examining the interaction effect of treatment and FSM status using the whole sample, following the same specification outlined above for EAL subgroup, and replicating the original primary outcome model on the FSM subgroup in a separate model.

Sensitivity analysis

Given the latent variables used in this evaluation differ to the one used in the previous evaluation (i.e., the CFA is from a different underlying SEM),, further sensitivity analysis is conducted to explore the relationship between the two latent variables. Both constructs can be measured at baseline and at the initial endline of the effectiveness trial. We replicate the primary outcome results presented in the effectiveness trial, using the new latent oral language variable instead, and compare these to the initial results to examine whether the results presented in the effectiveness trial are robust to latent variable specification. Should the new latent construct replicate the initial results, we can be more confident in being able to compare results at follow-up to results presented in the effectiveness trial despite the change in primary outcome construct.

Whilst EEF analysis guidelines generally recommend using pre–post analysis as a sensitivity analysis, we allow for language ability at initial endline to act as a moderator of ability at follow-up. This can help establish whether any effect size at follow-up is due to additional gains to NELI being realised in later primary years (in which case effect size should remain substantial even with the introduction of endline language ability) or due to the effects of NELI persisting in later primary years but without additional gains (in which case any positive effect size should largely disappear with the introduction of endline language ability). We include latent language ability at the effectiveness trial endline as an additional covariate in the X_{ij} in model (1), alongside baseline language.

Finally, we also explore the extent to which age at time of test impacts on the point estimate by including age at time of test as additional covariate in the model. This can help establish whether any effect size at follow-up is due to age or the effects of NELI.

Estimation of effect sizes

We used the effect size for cluster-randomised trials given in the EEF evaluators' guidance, adapted from Hedges (2007):

$$s^* = \frac{(\bar{Y}_T - \bar{Y}_C)_{adjusted}}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}}$$

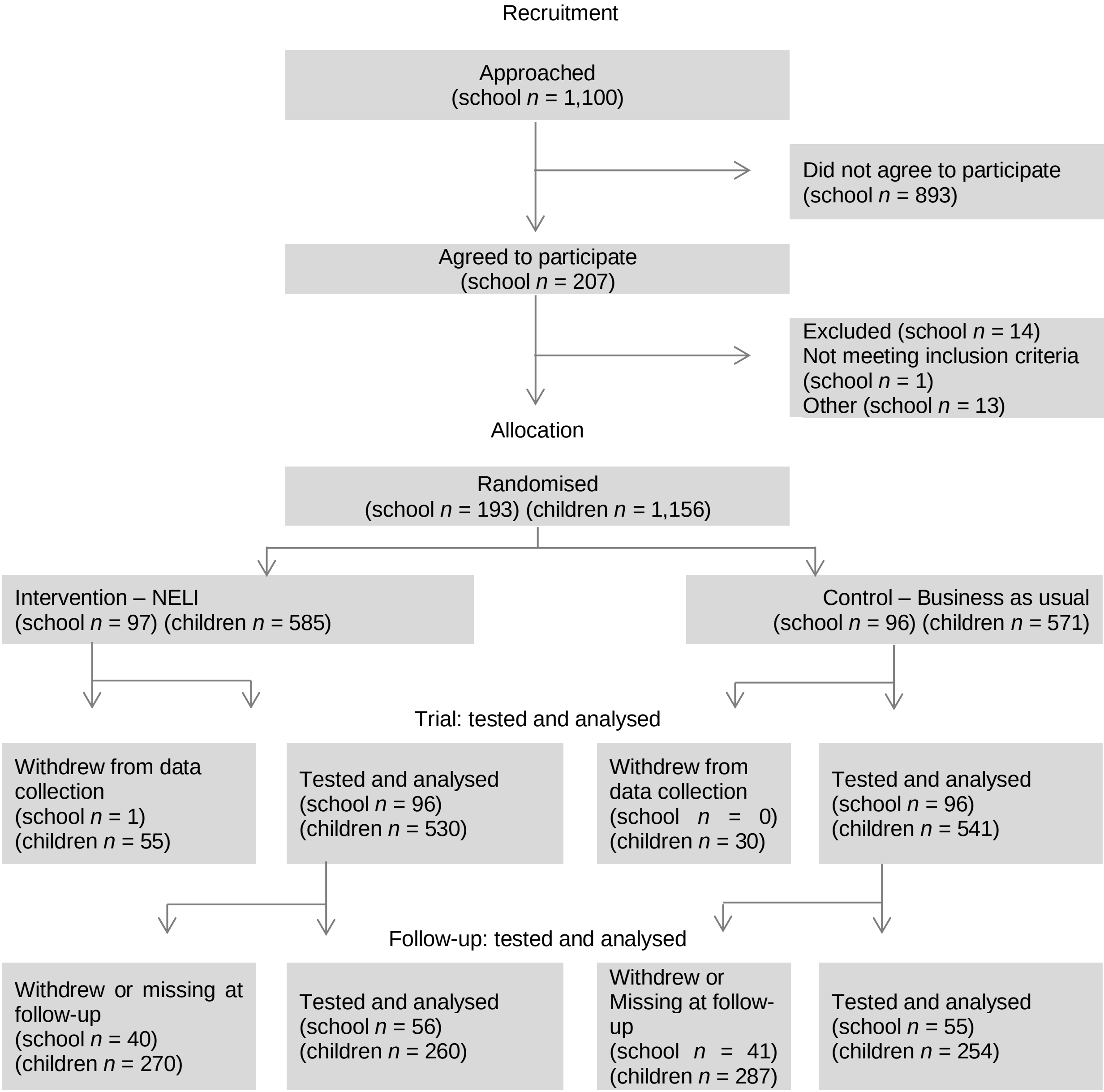
where s^* represents the clustered effect size, $(\bar{\boldsymbol{Y}}_{\boldsymbol{T}} - \bar{\boldsymbol{Y}}_{\boldsymbol{C}})_{\boldsymbol{adjusted}}$ is the mean difference between intervention groups adjusted for baseline characteristics (i.e., the coefficient from the multilevel model), and the denominator is an estimate of the pooled variance. A 95% confidence interval for the ES that takes into account the clustering of pupils in schools is reported.

Impact evaluation

Attrition

As can be seen in Figure 2, at randomisation there were 193 schools and an average of just over 6 pupils per school in the trial. At effectiveness trial analysis stage, 192 schools provided data, with an average of 5.7 children per school. At follow-up, 111 schools provided data, with an average of 4.6 pupils per school.

Figure 2: participant flow diagram


Recruitment
Approached (school *n* = 1,100)
Did not agree to participate (school *n* = 893)
Agreed to participate (school *n* = 207)
Excluded (school *n* = 14) Not meeting inclusion criteria (school *n* = 1) Other (school *n* = 13)
Allocation
Randomised (school *n* = 193) (children *n* = 1,156)
Intervention – NELI (school *n* = 97) (children *n* = 585)
Control – Business as usual (school *n* = 96) (children *n* = 571)
Trial: tested and analysed
Withdrew from data collection (school *n* = 1) (children *n* = 55)
Tested and analysed (school *n* = 96) (children *n* = 530)
Withdrew from data collection (school *n* = 0) (children *n* = 30)
Tested and analysed (school *n* = 96) (children *n* = 541)
Follow-up: tested and analysed
Withdrew or missing at follow-up (school *n* = 40) (children *n* = 270)
Tested and analysed (school *n* = 56) (children *n* = 260)
Withdrew or Missing at follow-up (school *n* = 41) (children *n* = 287)
Tested and analysed (school *n* = 55) (children *n* = 254)

Overall, attrition at the pupil level was 55.5% from randomisation to follow-up. While this seems to be fairly equal across treatment and control arms (see Table 3) it represents a significant amount of attrition. The high attrition is likely to be linked to the significant complications for schools and the evaluation in the face of the pandemic.

Table 3: Attrition

		Intervention	Control	Total

Number of pupils	Randomised	585	571	1156
	Analysed in main report	530	541	1071
	Analysed at follow-up	260	254	514
Pupil attrition (from randomisation to follow-up)	Number	325	317	642
	Percentage	55.6%	55.5%	55.5%

We explored attrition across trial arms as a basic step to assess bias (Higgins et al., 2011). As can be seen in Table 4, there have been some noticeable changes. The proportion of pupils eligible for FSM for schools at follow-up is lower than that at randomisation, suggesting that schools with more disadvantaged pupils were more likely to drop out at follow-up. The fall in the number of schools rated inadequate from randomisation to follow-up is likely accounted for by the rise in schools now in special measures in the case of control, and by the increase in schools which require improvement for the intervention. That intervention schools have improved performance, according to OFSTED rating, between randomisation and follow-up, whilst the number of schools on special measures has risen for control, may result in an upward bias in the estimated effect size.

At the pupil level, there is a fall in the number of EAL pupils at follow-up, when compared to at randomisation, suggesting the EAL subgroup analysis will be underpowered and results unlikely to be generalisable. Otherwise, there are no substantial differences in pupil-level data between randomisation and follow-up.

Table 4: Baseline characteristics of treatment and control

School-level (categorical)	**National -level mean**	**Full sample at randomisation**				**Follow-up sample only**			
		Intervention		**Control**		**Intervention**		**Control**	
		*n/N** (missing)	Count (%)	*n/N** (missing)	Count (%)	*n/N** (missing)	Count (%)	*n/N** (missing)	Count (%)
OFSTED rating									
1: Outstanding	12.06%	84/97 (11)	13.10%	75/96 (21)	18.67%	53/56 (3)	13.21%	47/56 (9)	12.77%
2: Good	62.35%	84/97 (11)	77.38%	75/96 (21)	64%	53/56 (3)	75.47%	47/56 (9)	74.47%
3: Requires Improvement	8.29%	84/97 (11)	4.76%	75/96 (21)	13.33%	53/56 (3)	11.32%	47/56 (9)	6.38%
4: Inadequate	0.01%	84/97 (11)	4.76%	75/96 (21)	4%	53/56 (3)	0%	47/56 (9)	0%
5: Special measures	1.44%	NA	NA	NA	NA	53/56 (3)	0%	47/56 (9)	6.38%

School proportion eligible for FSM	22.91%	87/97 (10)	34.05%	80/96 (16)	33.81%	55/56 (1)	23.98%	56/56 (0)	26.04%
School proportion classed as EAL	16.09%	NA	NA	NA	NA	55/56 (1)	12.92%	56/56 (0)	11.31%
School-level (continuous)		*n/N* (missing)	Mean (SD)	*n/N* (missing)	Mean (SD)	*n/N*(missing)	Mean (SD)	*n/N* (missing)	Mean (SD)
KS2 maths	104.94 (2.78)	85/97 (12)	104.12 (2.81)	80/96 (16)	103.71 (2.89)	48/56 (8)	104.69 (2.80)	46/56 (10)	104.17 (3.56)
KS2 reading	104.54 (2.92)	85/97 (12)	104.88 (2.93)	80/96 (16)	104.36 (2.80)	48/56 (8)	104.54 (2.94)	46/56 (10)	103.35 (3.24)
Pupil-level (categorical)		*n/N* (missing)	Count (%)	*n/N* (missing)	Count (%)	*n/N* (missing)	Count (%)	*n/N* (missing)	Count (%)
Gender									
Male		585/585	57.4%	571/571	53.2%	260/260	56.54%	254/254	52.76%
Female		585/585	42.6%	571/571	46.8%	260/260	43.46%	254/254	47.24%
EAL		537/585 (48)	33.33%	560/571 (11)	33.75%	260/260	28.46%	254/254	24.02%
FSM eligibility						260/260	23.85%	254/254	20.10%
Pupil-level (continuous)		*n/N* (missing)	Mean (SD)	*n/N* (missing)	Mean (SD)	*n/N* (missing)	Mean (SD)	*n/N* (missing)	Mean (SD)
Age (in months)		585/585 (0)	52.19 (3.49)	571/571 (0)	52.37 (3.51)	260/260 (0)	52.18 (3.54)	254/254 (0)	52.13 (3.41)
LanguageScreen listening comprehension		562/585 (23)	1.88 (2.01)	553/571 (18)	1.91 (1.91)	260/260 (0)	2.07 (2.04)	254/254 (0)	2.07 (1.89)
LanguageScreen expressive vocabulary		562/585 (23)	7.24 (4.17)	553/571 (18)	7.25 (3.86)	260/260 (0)	7.64 (4.10)	254/254 (0)	8.06 (3.72)
LanguageScreen receptive vocabulary		562/585 (23)	15.65 (4.16)	553/571 (18)	15.74 (4.13)	260/260 (0)	16.05 (4.14)	254/254 (0)	16.25 (4.14)
LanguageScreen sentence repetition		562/585 (23)	2.41 (2.47)	553/571 (18)	2.44 (2.47)	260/260 (0)	2.48 (2.38)	254/254 (0)	2.56 (2.43)
RAPT information		585/585 (0)	19.11 (7.86)	585/585 (0)	20.09 (7.36)	260/260 (0)	19.87 (7.99)	254/254 (0)	21.19 (6.95)
RAPT grammar		585/585 (0)	11.69 (6.87)	571/571 (0)	12.11 (6.57)	260/260 (0)	12.21 (6.80)	254/254 (0)	12.59 (6.31)
YARC early word reading		585/585 (0)	0.61 (2.99)	571/571 (0)	0.48 (2.37)	256/260 (4)	25.05 (0.52)	251/254 (3)	23.86 (0.44)

Despite these possible differences at the school level, our analyses of missingness (i.e., assessing whether there are systematic differences between those who drop out and those who do not) show no significant differences by allocation to treatment or by any of the above predictors. The observed variables are not significant predictors of missingness (Appendix B). We proceed with a complete case analysis of data at follow-up, and instead recommend caution with interpretation of the primary and secondary outcome analysis.

As observed variables are not significant predictors of missingness, attrition seems likely to be related to other unobservable factors, such as differential impact of Covid on different schools and/or pupils, which may negatively affect pupil attainment. For instance, schools in areas that were hit harder by Covid lockdowns during the course of 2020 or had Covid outbreaks in the Autumn of 2021 when testing was underway, may be more likely to be missing in this dataset, as their priority may have shifted towards making up for lost time due to Covid rather than undertaking testing. Such attrition is likely to be negatively correlated with achievement, with prolonged school closure or disruption thought to act as a barrier to learning. This hypothesised negative relationship between missingness and pupil progression would suggest that any point estimates here are an upper bound. However, as treatment was randomised, it seems unlikely that this would disproportionately effect treatment over control: so long as any unobserved variables affect the two groups equally, the effect size may still be unbiased.

The high level of attrition presents a threat to external validity of the estimates. Our analysis indicates that attrition is largely due to unobservable factors, which is problematic for the interpretation of any effect sizes; if missing schools and pupils differ from those who remain at follow-up, results are not necessarily generalisable to the wider population.

Analysis

All variables used in analysis are summarised in Table 5.

Table 5: Summary statistics for variables used in analysis

	***n*/*N* (missing)**	**Mean**	**SD**
Latent oral language	507/514 (7)	0.011	2.03
LanguageScreen listening comprehension	511/514 (3)	11.15	2.82
LanguageScreen expressive vocabulary	511/514 (3)	15.68	3.56
LanguageScreen receptive vocabulary	511/514 (3)	19.21	2.71
LanguageScreen sentence repetition	511/514 (3)	10.44	2.58
RAPT information	510/514 (4)	28.73	4.56
RAPT grammar	510/514 (4)	24.93	5.48
YARC early word reading	510/514 (4)	24.47	7.67
YARC reading fluency	509/514 (5)	14.50	17.72
YARC reading comprehension	509/514 (5)	13.44	6.70

Reading outcome analysis

Analysis was carried out as per the *Methods* section, again using a multilevel model to account for the clustering of pupils in schools and for the pre-test under an intention-to-treat approach. All three outcome variables are non-normal as can be clearly seen in the histograms inAppendix A. After running the multilevel model, the normality of residuals was examined for all three outcome variables, and found to be non-normal (see Appendix A). Whilst point estimates are likely still unbiased in the presence of non-normality, standard errors and confidence intervals are unlikely to be correctly estimated using standard approaches. To overcome the non-normality assumption, we used the EEF Analytics package in Stata to generate bootstrapped Hedge's *g* and confidence intervals. Results for reading outcomes are reported in **Error! Reference source not found.**Tables 7 and 8, with details of the SEM reported in Appendix C**Error! Reference source not found.**.

Table 6: Reading outcome analysis results

			Intervention group		Control group		
Outcome	Unadjusted differences in means	Adjusted differences in means	*n* (missing)	Variance of outcome	*n* (missing)	Variance of outcome	Pooled variance
Early word reading	1.01	1.37	260 (0)	52.21	254 (0)	67.44	59.74
Reading fluency	1.36	0.78	259 (1)	368.6	254 (0)	260.6	315.13
Reading comprehension	0.675	1.162	259 (1)	44.93	254 (0)	45.32	45.123

Table 7: Effect size estimation – reading outcome

	Raw means				Effect size	
	Intervention group		Control group			
Outcome	*n* (missing)	Mean (95% CI)	*n* (missing)	Mean (95% CI)	*n* in model (intervention; control)	Hedges *g* (95% CI)
Early word reading	260 (0)	24.92 (24.04, 25.80)	254 (0)	23.91 (22.89, 24.92)	514 (260; 254)	0.18 (0.02, 0.39)
Reading fluency	259 (1)	15.14 (12.79, 17.49)	254 (0)	13.78 (11.79, 15.78)	513 (259, 254)	0.04 (−0.14, 0.20)
Reading comprehension	259 (1)	13.73 (12.91, 14.55)	254 (0)	13.05 (12.22, 13.89)	513 (259, 254)	0.17 (0.00, 0.35)

Both early word reading and reading comprehension had effect sizes in line with the language outcome (0.179 and 0.173, respectively, see *Language outcome analysis* section below). This suggests two key points for the logic model (Figure 1). Firstly, that NELI does appear to be linked to improved reading comprehension and could be considered an appropriate longer term outcome. Secondly, that the impact of NELI on early word reading is sustained to a certain extent over time for those pupils involved in the follow-up evaluation. However, it is worth noting that there were ceiling effects in early word reading. Whilst primary-age YARC has been validated for use in children aged 4–11, the different tests are suited to different age groups: Early Word Reading (EWR) is specifically designed for younger pupils (mostly those in EYFS or early stages of KS1). In theory, most children should have 'aged out' of the early word reading which may explain the ceiling effects. The presence of substantial ceiling effects suggests any estimates should be interpreted with caution.

In contrast, the effect size for reading fluency (0.043) is lower than other outcomes. This suggests that NELI has less of an impact on reading fluency for those pupils involved in the follow-up evaluation. However, it is worth noting that floor effects were found. The floor effects on reading fluency make it difficult to explain the null effect found. It cannot be ascertained whether the apparent lack of effect on reading fluency is due to a true null treatment effect or due to the test's lack of sensitivity to small improvements. Reading comprehension still exhibits some floor effects, but not the same degree as with reading fluency. This suggests that the YARC reading comprehension test was more sensitive than the other measures to differences in reading ability with this age and ability range.

Language outcome analysis

Details of the language outcome model are reported in Table 8 and Table 9. At follow-up, the raw mean for the intervention group was larger than the raw mean for the business-as-usual control group and the effect size associated with this difference is 0.18. The conclusion of the language outcome analysis is, therefore, that NELI continues to have a sustained influence on oral language for those pupils included in the follow-up analysis. This conclusion is further supported by the sensitivity analysis (discussed below), which suggests that impact is sustained to a certain extent over time. However, given the high rate of attrition in the sample, this result should be interpreted with caution and generalisation of the results avoided.

Table 8: Language outcome analysis results

			Intervention group		Control group		
Outcome	Unadjusted differences in means (I-C)	Adjusted differences in means	*n* (missing)	Variance of outcome	*n* (missing)	Variance of outcome	Pooled variance
Language	0.170	0.368	259 (1)	4.393	248 (6)	3.860	4.132

Table 9: Effect size estimation – language outcome

	Raw means				Effect size	
	Intervention group		Control group			
Outcome	*n* (missing)	Mean (95% CI)	*n* (missing)	Mean (95% CI)	*n* in model (intervention; control)	Hedges *g* (95% CI)
Language	259 (1)	0.094 (−0.163, 0.350)	248 (6)	−0.076 (−0.322, 0.170)	507 (259; 248)	0.181 (0.02, 0.35)

Sensitivity analysis

Table 10 and Table 11 outline the results from the sensitivity analysis, that looks at the extent to which impact was, in fact, sustained from effectiveness trial to follow-up for those measures that were collected across both evaluations. Results show that the effect we observe at follow-up is due to persistence of treatment effect rather than any additional gains made over time, as the effect disappears once we control for results at the endline of the effectiveness trial. This provides us with increased confidence that the impact of NELI was in fact sustained to some extent. Promisingly, these sustained effects seem to be transferred to less proximal outcomes, such as reading comprehension.

Table 10: Language outcome analysis results – including latent language at initial endline as a covariate in model (1)

			Intervention group		Control group		
Outcome	Unadjusted differences in means (I-C)	Adjusted differences in means	*n* (missing)	Variance of outcome	*n* (missing)	Variance of outcome	Pooled variance
Language sustained	0.093	0.065	229 (31)	4.556	213 (41)	3.693	4.140

Table 11: Effect size estimation: primary outcome when including latent language at initial endline as a covariate in model (1)

	Raw means				Effect size	
	Intervention group		Control group			
Outcome	*n* (missing)	Mean (95% CI)	*n* (missing)	Mean (95% CI)	*n* in model (intervention; control)	Hedges *g* (95% CI)
Language sustained	229 (31)	0.052 (−0.226, 0.330)	213 (41)	−0.040 (−0.300, 0.219)	442 (229; 213)	0.032 (−0.12, 0.19)

Table 12 and Table 134 outline the results from the sensitivity analysis, to understand whether the latent variable used at follow-up is comparable to the latent variable used for the effectiveness trial.

Table 12: Language outcome analysis results – at effectiveness trial stage, using new latent construct

			Intervention group		Control group		
Outcome	Unadjusted differences in means (I-C)	Adjusted differences in means	*n*	Variance of outcome	*n*	Variance of outcome	Pooled variance
Language	0.713	1.18	434	15.422	433	17.304	16.362

Table 13: Effect size estimation – at effectiveness trial stage, using new latent construct

	Raw means				Effect size	
	Intervention group		Control group			
Outcome	*n*	Mean (95% CI)	*n*	Mean (95% CI)	*n* in model (intervention; control)	Hedges *g* (95% CI)
Language	434	0.365 (−0.006, 0.735)	433	−0.348 (−0.741, 0.045)	867 (434; 433)	0.292 (0.19, 0.40)

The original estimate treatment effect for the language outcome can be replicated to within three-hundredths using the new latent oral language factor in place of the original latent oral language factor. These results suggest that this new construct measures the same underlying construct as the original model used in the effectiveness trial. This provides us with confidence that the impact seen in this follow-up study are a result of sustained impact of NELI and not due to a change in how the measure was constructed.

The correlation between the original latent factor and the new latent factor for language is 0.80 at endline and 0.91 at baseline, suggesting that the two latent factors are measuring a similar concept, or at least one which correlates well with the original model. Given there is an overlap in the included measures for both models, however, it is unsurprising the two latent factors are highly correlated.

Finally, given testing at follow-up was conducted over two terms (summer and autumn terms of 2021), we additionally control for age in months at follow-up as a sensitivity analysis. There is no clear evidence that treatment effect is age-dependent, suggesting that collecting over two terms does not appear to have significantly biased the results. The output for this sensitivity analysis is provided in Appendix D.

Subgroup outcome analysis

Details of the EAL subgroup model are reported in

Table 14 and Table 15. The analysis indicates that NELI had a positive effect on EAL pupils' language skill development in the intervention group compared to the control group. However, given the high rate of attrition in the sample, and the relatively low numbers of EAL pupils in this analysis, these subgroup results should be interpreted with caution.

Table 14: Language outcome subgroup analysis results – English as an additional language (EAL), when included as an interaction term in model (1)

			Intervention group		Control group		
Outcome	Unadjusted differences in means (I-C)	Adjusted differences in means	*n* (missing)	Variance of outcome	*n* (missing)	Variance of outcome	Pooled variance
Primary outcome: language	−0.04	0.528	74 (0)	6.056	60 (0)	3.670	4.990

Table 15: Effect size estimation – language outcome for EAL subgroup (when included as an interaction term) in model (1)

	Raw means		Effect size
	Intervention group	Control group	

Outcome	*n* (missing)	Mean (95% CI)	*n* (missing)	Mean (95% CI)	*n* in model (intervention; control)	Hedges *g* (95% CI)
Primary outcome: language	74 (0)	−0.653 (−1.223, 0.083)	60 (0)	−0.693 (−1.188, −0.198)	134 (74; 60)	0.236 (−0.04, 0.51)

Details of the FSM subgroup model are reported in Tables 17 and 18. It is estimated that NELI has a positive additional effect on FSM pupils, as measured by a treatment effect in the original model. However, this analysis is underpowered and so the estimated effect size is not found to be significant. As FSM was not analysed in the original evaluation report, however, we undertake further exploratory analysis by calculating the effect size of NELI on the FSM subgroup, reported in Tables 17 and 18. In magnitude, the estimated effect size of NELI on the FSM subgroup is larger than that found for wider sample, suggesting NELI may act as a gap closer. However, such an interpretation is limited by sample size and should be seen as exploratory.[3]

Table 16: Language outcome subgroup analysis results – free school meals (FSM), when included as an interaction term in model (1)

			Intervention group		**Control group**		
Outcome	Unadjusted differences in means (I-C)	Adjusted differences in means	*n* (missing)	Variance of outcome	*n* (missing)	Variance of outcome	Pooled variance
Language	−0.164	0.418	51 (0)	6.664	62 (0)	3.521	4.937

[3] For instance, low power increases the risk that the effect size may have an incorrect sign (Gelman and Carlin, 2014).

Table 17: Effect size estimation – language outcome for FSM pupils, when included as an interaction term in Model (1)

	Raw means				**Effect size**	
	Intervention group		**Control group**			
Outcome	*n* (missing)	Mean (95% CI)	*n* (missing)	Mean (95% CI)	*n* in model (intervention; control)	Hedges *g* (95% CI)
Language	51 (0)	−0.618 (−1.344, 0.108)	62 (0)	−0.782 (−1.260, −0.306)	113 (51; 62)	0.188 (−0.08, 0.46)

Table 18: Language outcome subgroup analysis results – free school meals (FSM) subgroup

			Intervention group		**Control group**		
Outcome	Unadjusted differences in means (I-C)	Adjusted differences in means	*n* (missing)	Variance of outcome	*n* (missing)	Variance of outcome	Pooled variance
Language	−0.164	0.537	51 (0)	6.664	62 (0)	3.521	4.937

Table 19: Effect size estimation – language outcome for FSM subgroup

	Raw means				**Effect size**	
	Intervention group		**Control group**			
Outcome	*n* (missing)	Mean (95% CI)	*n* (missing)	Mean (95% CI)	*n* in model (intervention; control)	Hedges *g* (95% CI)
Language	51 (0)	−0.618 (−1.344, 0.108)	62 (0)	−0.782 (−1.260, −0.306)	113 (51; 62)	0.242 (−0.14, 1.22)

Analysis in both EAL and FSM subgroups was limited to the proximal outcome, due to the sample size restrictions. As NELI is targeted towards oral language, we might expect the impact on oral language to be stronger than any long-term outcomes, so if we are to find any effect, we would expect to see it first in oral language. The analysis above is indicative of a positive effect,[4] but due to small sample size, the analysis is unable to detect an effect for the proximal outcome. Further analysis of potential long-run outcomes, such as reading comprehension, would be similarly limited and underpowered.

Conclusion

Key conclusions
Much of the gains to oral language as a result of NELI persist over time for those pupils included in the follow-up analysis.
The effect of NELI on early word reading persists over time for those pupils included in the follow-up analysis.
NELI appears to lead to improved reading comprehension for those pupils included in the follow-up analysis. This is in line with the logic model with improvements in language leading to improvements in reading comprehension.
There appears to be a larger effect of NELI on language for FSM-eligible pupils compared to non-FSM-eligible pupils. However, this finding should be interpreted with caution given the low statistical power.

Interpretation

This follow-up report presents the findings of the largest trial of NELI ever undertaken, approximately two years after it was implemented. The aim of this follow-up study was threefold. Firstly, to explore the extent to which the impact

[4] As commented previously, a lack of statistical power increases the risk errors in sign as well as magnitude when estimating the effect size.

seen on language and early reading in the NELI effectiveness trial transfer to reading comprehension and fluency. Secondly, to explore the extent to which language differences between pupils who took part in NELI and 'business-as-usual' pupils persist. Finally, to explore whether there are differential effects of NELI on the oral language of FSM-eligible pupils compared to non-FSM-eligible pupils.

Our exploratory analysis on reading outcomes suggests that there is a transfer from language to reading comprehension for those pupils involved in the follow-up evaluation, in line with the hypothesised logic model. In addition, the exploratory analysis also suggests that the impact on early word reading is sustained. Finally, there is limited evidence of NELI on pupils' reading fluency for those pupils that provided data for the follow-up. This could be because NELI has no impact on that domain – which would be in line with the logic model generated in the effectiveness trial. Alternatively, it could be that the measure of fluency was not sensitive enough to detect effects, which is supported by the floor effects seen in the measure (see

Figure 8 in Appendix A).

Our analysis of language outcomes suggests that the impact of NELI on pupils' language skills was sustained over time for those pupils that provided data for the follow-up. This suggests that NELI is a potentially beneficial intervention that has the ability to support the spoken language ability of children in primary school who have spoken language skills that are poorer than those of their peers, with results that are sustained in the long term.

In the exploratory subgroup analyses, we find the effect of NELI on language is larger for FSM-eligible pupils than it is for non-FSM-eligible pupils, for those pupils included in the follow-up analysis. There is also a sustained effect of NELI on the language skills of pupils who have EAL compared to EAL pupils in the control group. This result should be interpreted with caution, given the small sample size and the high attrition rates seen in the follow-up evaluation.

These findings should be interpreted with caution given their exploratory nature, as the original effectiveness trial was not explicitly designed to test these research questions. High attrition and relatively small sample size add to the cautionary nature of any interpretation (see discussion below).

Limitations

The key limitation of this study is the attrition, with 55.5% of pupils missing at follow-up from endline. This could be linked to a number of factors – the fact that schools were focusing on catch-up post-Covid leaving them little time to support testing, natural attrition in the sample, and pupil turnover. There were some noticeable differences observed between schools in treatment and control, with regard to OFSTED rating and proportion of pupils eligible for FSM, that increase the risk of introducing bias and reduce generalisability of results. Our analysis of attrition, however, suggested that there were no significant predictors of missingness, suggesting that there were unobservable factors that led to missing data. This makes interpretation of results difficult, as there is no characteristic at school or pupil level that allows us to make any generalisations (e.g., all Good schools remained in the trial, all EAL pupils were missing at follow-up). Attrition is likely to be related to unobservable factors, such as the differential impact of Covid on different schools and/or pupils, with those schools who faced more prolonged lockdowns or further waves in 2021 both less likely to participate in testing at follow-up and more likely to find their pupils are lagging behind where they would normally be, due to disruption to time in school. Given this hypothesised negative relationship between missingness and pupil progression, it is possible that point estimates provided here are an upper bound. However, randomisation at baseline ensures that Covid and other exogenous shocks would be expected to effect treatment and control equally, and so the estimated effect size may remain unbiased.

School closures and limitations on school visits as a result of Covid meant that the language tests used in the original effectiveness trial could not be used at follow-up. As a result, the latent variable used to measure language is different in the follow-up study compared to the effectiveness trial (see Table 2**Error! Reference source not found.**). A sensitivity analysis suggests that the latent variable of language used in this report mirrors the findings from the effectiveness trial (see Table 13), which increases our confidence that the new latent variable is measuring a similar construct to that in the effectiveness trial. However, the fact that the follow-up study does not include the same measures is a limitation.

A further limitation is the fact that the LanguageScreen measure, developed by the delivery team, became a key part of the primary measure – with four out of six components of the final latent variable being made up of LanguageScreen tests. One potential criticism is that the measure was developed by the delivery team, which could introduce bias as the measure may be aligned to the intervention (i.e., by asking specific content questions covered by the NELI programme). However, the measure was designed to be a general screening tool, not specifically or solely for NELI, and given the practical requirements, this was the most pragmatic option. The robustness checks conducted on this new model suggest that, by and large, the two models are highly correlated and produce effect sizes of a similar magnitude, suggesting that, in practice, the inclusion of the LanguageScreen variables has not substantially biased the results.

Analysis of reading outcomes was limited by the presence of substantial ceiling and floor effects. Many of the children had aged out of EWR at follow-up, generating ceiling effects, but YARC reading fluency, and to a lesser extent reading comprehension, were still not sensitive enough over this age and ability range. The presence of substantial ceiling and floor effects for all three YARC measures suggests any estimates should be interpreted with caution.

Further research

Future research would do well to focus on understanding how NELI impacts on learning over the longer term. An analysis of children's outcomes at KS2 would be particularly interesting, to see how children who took place in NELI compare to children in control schools that did not. The findings from this exploratory analysis suggest that language is linked to reading so looking at reading outcomes at KS2 would be interesting to follow. Research also suggests that language is linked to performance in maths (Purpura et al., 2019) so it would be interesting to look at pupils' maths results at KS2. This would be possible given data for the main trial have been archived and are housed in the EEF data archive, including Unique Pupil Numbers (UPNs).

Finally, it is worth noting that the Department of Education (DfE) has been working with the EEF and other delivery partners to make NELI available to state-funded primary schools with online training and resources being made available at no cost for schools. This is part of a wider £1 billion of funding announced on 19 June 2020 to help pupils and disadvantaged young people catch up on missed education because of coronavirus (Covid-19).

References

Acock, A. C. (2013) Discovering structural equation modeling using Stata. Stata Press.

Bowyer-Crane, C., Fricke, S., Schaefer, B., Lervåg, A. & Hulme, C. (2017) ‘Early literacy and comprehension skills in children learning English as an additional language and monolingual children with language weaknesses’, *Reading and Writing*, 30(4), 771–790.

Dimova, S., Ilie, S., Brown, E.R., Broeks, M., Culora, A., & Sutherland, A. (2020) ‘Nuffield Early Language Intervention: Evaluation Report’, Education Endowment Foundation.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6), 641–651.

Heß, S. (2017) ‘Randomization Inference with Stata: A guide and software’, *Stata Journal*, 17(3), 630–651.

Higgins, J. P., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., Savović, J., Schulz, K. F., Weeks, L. & Sterne, J. A. (2011) ‘The Cochrane Collaboration’s Tool for assessing risk of bias in randomised trials’, *BMJ*, 343, d5928.

Hox, J. (1998) ‘Multilevel modeling: When and why’, in *Classification, Data Analysis, and Data Highways*, Berlin: Springer.

Hu Li-tze & Bentler P. M. (1999) Cut-off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives, *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55.

Hulme, C., Duta, M., West, G., Cripps, D., Dasgupta, A., Hearne, S., Gardner, R., & Snowling, M., (pre-print), LanguageScreen: The development and standardisation of a language assessment app, Submitted to: *Language Speech and Hearing Services in Schools*

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(1), 1–10.

Kenny, D. A. (2015) ‘Measuring model fit’: retrieved from http://www.davidakenny.net/cm/fit.htm

Lai, K. & Green, S. B. (2016) ‘The problem with having two watches: assessment of fit when RMSEA and CFI disagree’, *Multivariate Behavioral Research*, 51(2–3), 220–239.

Muthén, B.O. (2002) ‘Beyond SEM: general latent variable modeling’, *Behaviormetrika*, 29(1), 81–117.

Purpura, D. J., Napoli, A. R., & King, Y. (2019). ‘Development of mathematical language in preschool and its role in learning numeracy skills’, in D. C. Geary, D. B. Berch, & K. M. Koepke (Eds.), *Cognitive Foundations for Improving Mathematical Learning*, pp. 175–193. Elsevier Academic Press.

Schumacker, R. E. & Lomax, R.G. (2016) *A Beginner’s Guide to Structural Equation Modeling*, Fourth Edition. New York, NY: Routledge.

Shi, D., Lee, T., & Maydeu-Olivares, A. (2019) ‘Understanding the model size effect on SEM fit indices’, *Educational and Psychological Measurement*, 79(2), 310–334.

Snijders, T. A. B. (2005) ‘Power and sample size in multilevel modeling’, in B. S. Everitt and D. C. Howell (eds), *Encyclopedia of Statistics in Behavioral Science*, Chicester: Wiley (vol. 3, pp.1570–1573).

Snijders, T. A. & Bosker, R. J. (1994) ‘Modeled variance in two-level models’, *Sociological Methods and Research*, 22(3), p.342–363.

Appendix A: Histograms and Q–Q plots

Figure 3: Distribution of language outcomes at baseline

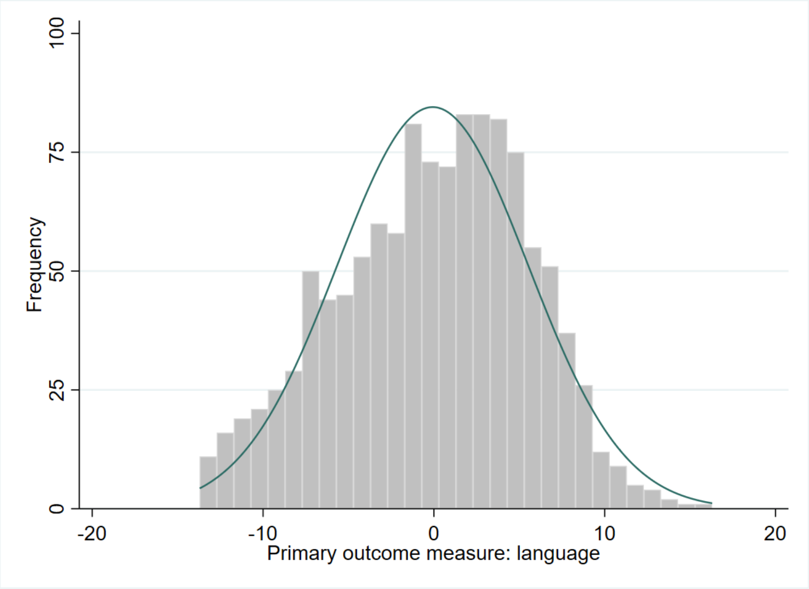


Figure 4: Distribution of language outcomes at the initial endline

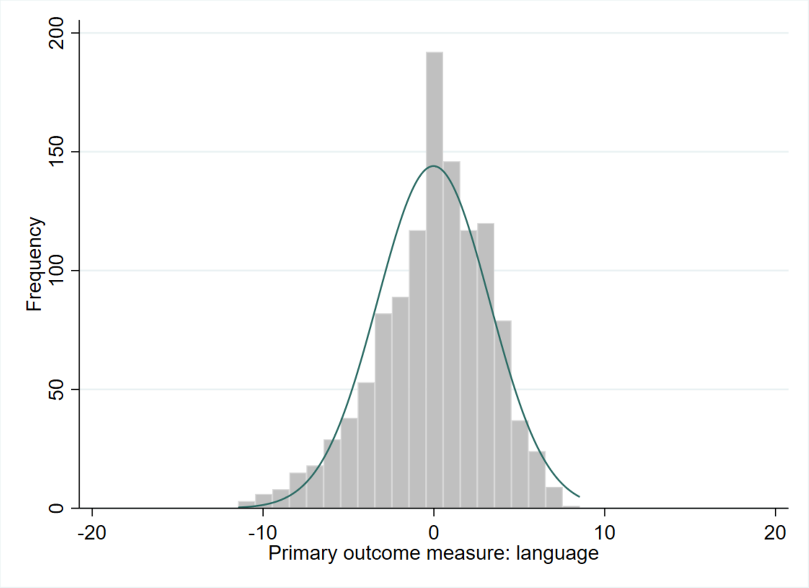


Figure 5: Distribution of language outcomes at follow-up

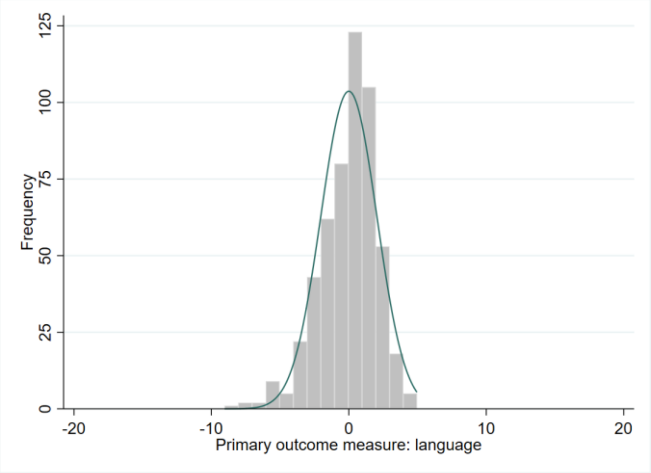


Figure 6: Distribution of YARC early word reading at follow-up

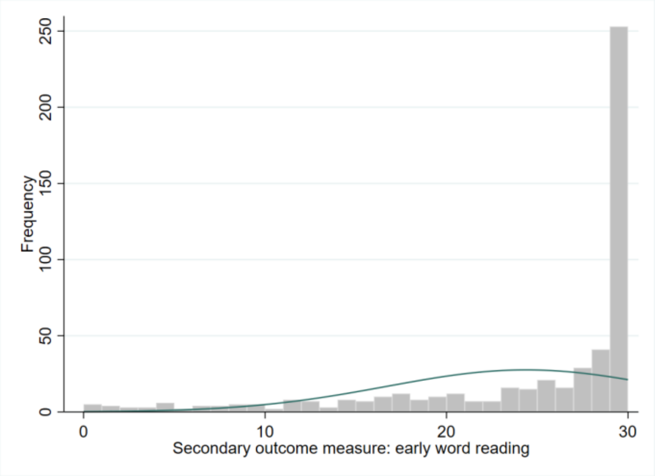


Figure 7: Distribution of YARC reading comprehension at follow-up

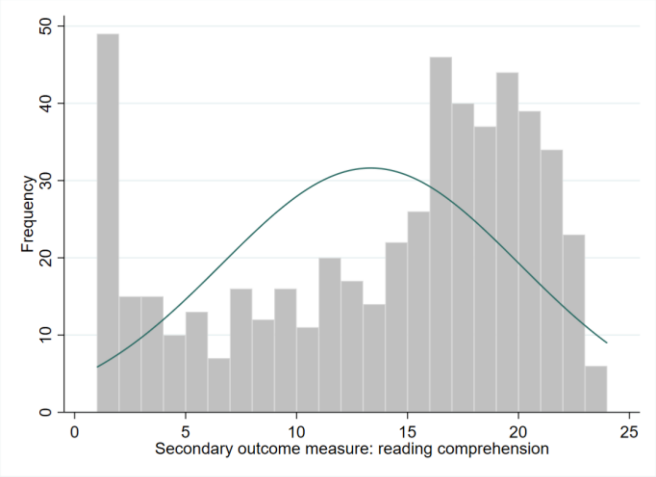


Figure 8: Distribution of YARC reading fluency at follow-up

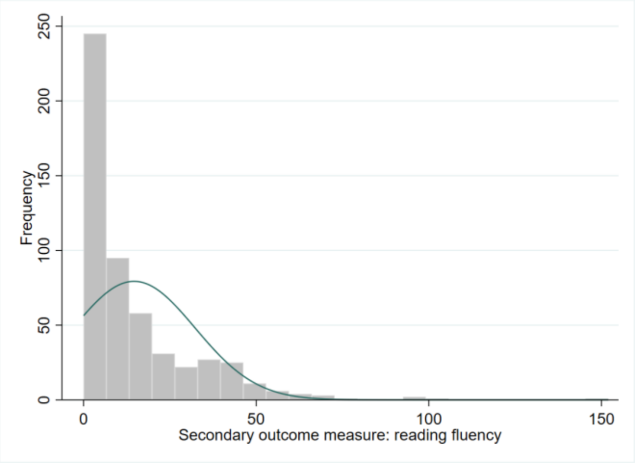


Figure 9: Q–Q plot of the residuals from the mixed effects model for YARC early word reading

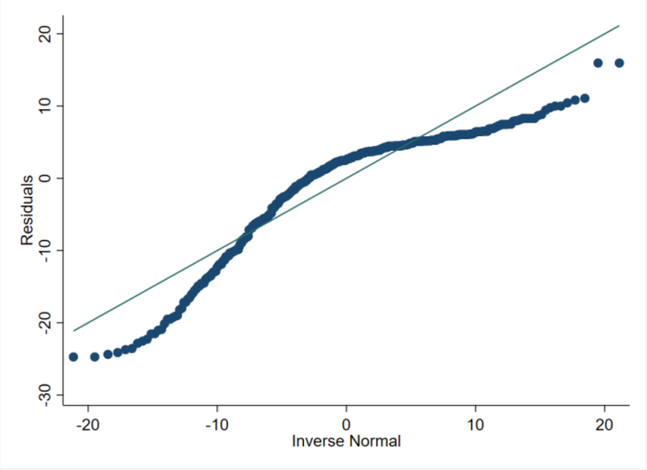


Figure 10: Q–Q plot of the residuals from the mixed effects model for YARC reading fluency

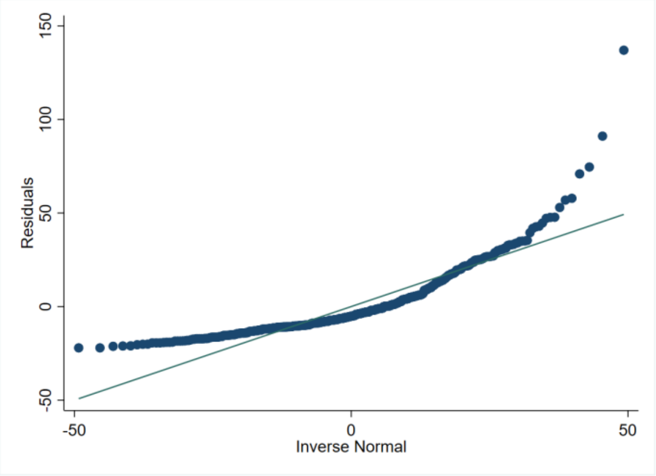
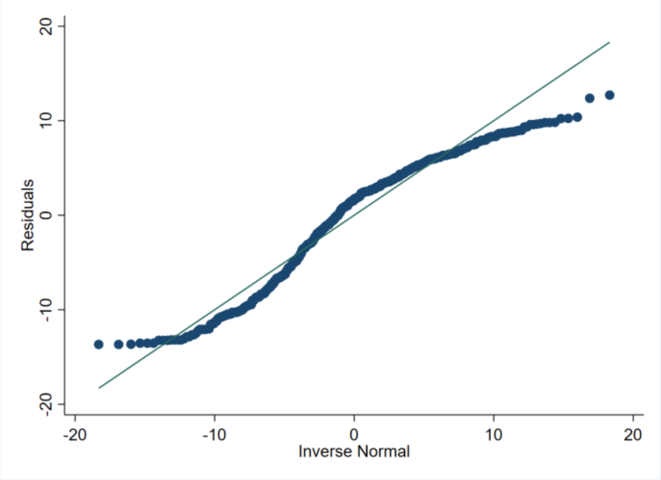
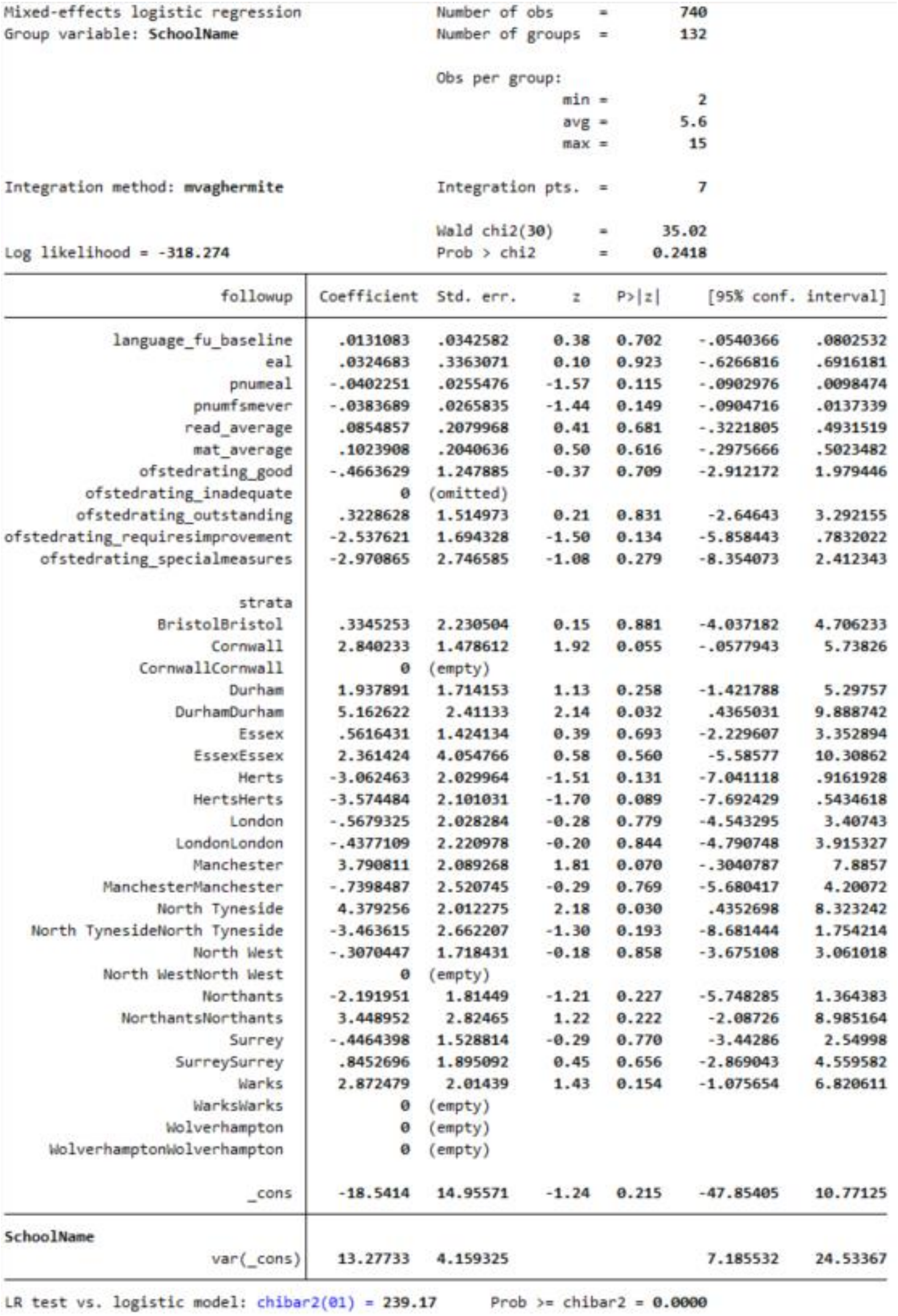


Figure 11: Q–Q plot of the residuals from the mixed effects model for YARC reading comprehension



Appendix B: Results from the Logit model constructed to examine missingness



```
Mixed-effects logistic regression               Number of obs     =        740
Group variable: SchoolName                      Number of groups  =        132

                                                Obs per group:
                                                              min =          2
                                                              avg =        5.6
                                                              max =         15

Integration method: mvaghermite                 Integration pts.  =          7

                                                Wald chi2(30)     =      35.02
Log likelihood = -318.274                       Prob > chi2       =     0.2418

------------------------------------------------------------------------------------
                         followup | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
----------------------------------+-------------------------------------------------------------------
             language_fu_baseline |   .0131083   .0342582     0.38   0.702    -.0540366    .0802532
                              eal |   .0324683   .3363071     0.10   0.923    -.6266816    .6916181
                          pnumeal |  -.0402251   .0255476    -1.57   0.115    -.0902976    .0098474
                      pnumfsmever |  -.0383689   .0265835    -1.44   0.149    -.0904716    .0137339
                     read_average |   .0854857   .2079968     0.41   0.681    -.3221805    .4931519
                      mat_average |   .1023908   .2040636     0.50   0.616    -.2975666    .5023482
                ofstedrating_good |  -.4663629   1.247885    -0.37   0.709    -2.912172    1.979446
          ofstedrating_inadequate |          0  (omitted)
        ofstedrating_outstanding |   .3228628   1.514973     0.21   0.831     -2.64643    3.292155
ofstedrating_requiresimprovement |  -2.537621   1.694328    -1.50   0.134    -5.858443    .7832022
    ofstedrating_specialmeasures |  -2.970865   2.746585    -1.08   0.279    -8.354073    2.412343
                                  |
                           strata |
                   BristolBristol |   .3345253   2.230504     0.15   0.881    -4.037182    4.706233
                         Cornwall |   2.840233   1.478612     1.92   0.055    -.0577943     5.73826
                  CornwallCornwall |          0  (empty)
                           Durham |   1.937891   1.714153     1.13   0.258    -1.421788     5.29757
                     DurhamDurham |   5.162622    2.41133     2.14   0.032     .4365031    9.888742
                            Essex |   .5616431   1.424134     0.39   0.693    -2.229607    3.352894
                       EssexEssex |   2.361424   4.054766     0.58   0.560     -5.58577    10.30862
                            Herts |  -3.062463   2.029964    -1.51   0.131    -7.041118    .9161928
                       HertsHerts |  -3.574484   2.101031    -1.70   0.089    -7.692429    .5434618
                           London |  -.5679325   2.028284    -0.28   0.779    -4.543295     3.40743
                     LondonLondon |  -.4377109   2.220978    -0.20   0.844    -4.790748    3.915327
                       Manchester |   3.790811   2.089268     1.81   0.070    -.3040787      7.8857
           ManchesterManchester |  -.7398487   2.520745    -0.29   0.769    -5.680417     4.20072
                   North Tyneside |   4.379256   2.012275     2.18   0.030     .4352698    8.323242
     North TynesideNorth Tyneside |  -3.463615   2.662207    -1.30   0.193    -8.681444    1.754214
                       North West |  -.3070447   1.718431    -0.18   0.858    -3.675108    3.061018
             North WestNorth West |          0  (empty)
                        Northants |  -2.191951    1.81449    -1.21   0.227    -5.748285    1.364383
               NorthantsNorthants |   3.448952    2.82465     1.22   0.222     -2.08726    8.985164
                           Surrey |  -.4464398   1.528814    -0.29   0.770     -3.44286     2.54998
                     SurreySurrey |   .8452696   1.895092     0.45   0.656    -2.869043    4.559582
                            Warks |   2.872479    2.01439     1.43   0.154    -1.075654    6.820611
                       WarksWarks |          0  (empty)
                     Wolverhampton |          0  (empty)
       WolverhamptonWolverhampton |          0  (empty)
                                  |
                            _cons |   -18.5414   14.95571    -1.24   0.215    -47.85405    10.77125
----------------------------------+-------------------------------------------------------------------
SchoolName                        |
                        var(_cons)|   13.27733   4.159325                      7.185532    24.53367
------------------------------------------------------------------------------------
LR test vs. logistic model: chibar2(01) = 239.17        Prob >= chibar2 = 0.0000
```

Appendix C: SEM model

Primary outcome measure

At follow-up, the original **SEM** model to derive the main oral language skills outcome measure was:

```
sem (LANG_fu -> t3APT_info t3APT_gram t3Evtotal t3Rvtotal t3Srtotal t3Lctotal), latent (LANG_fu)
standardized cov(e.t3APT_info*e.t3APT_gram)
```

The errors of the RAPT information and RAPT grammar variables were allowed to correlate with each other because during the test these two items are asked together. This was also consistent with the analysis of the main trial.

The initial model fit was acceptable but not good, as far as RMSEA is concerned: CFI = 0.976; RMSEA = 0.068.

The model was therefore modified, this time allowing the errors of the RAPT grammar and LanguageScreen sentence repetition variables to correlate with each other, as accurate sentence repetition was hypothesised to largely rely on the learner’s knowledge of grammar:

```
sem (LANG_fu -> t3APT_info t3APT_gram t3Evtotal t3Rvtotal t3Srtotal t3Lctotal), latent (LANG_fu)
standardized cov(e.t3APT_info*e.t3APT_gram) cov(e.t3APT_gram*e.t3Srtotal)
```

Both indices improved following the modification, indicating a good fit between the hypothesised model and the observed data: CFI = 0.99; RMSEA = 0.048. We then tested for a significant difference between the two models to decide which of the two to keep. At *p* = 0.0005, the modified model appeared significantly improved over the original one and was therefore retained.

A factor score for the primary outcome language variable was then estimated.

For all pupils with outcome data (537), this measure had an overall mean of 0 and standard deviation of 2.03 for the whole sample of children with outcome data, and a range from −9.04 to 4.32. This variable constitutes the primary outcome measure in this trial.

Secondary outcome measure

The YARC tests was used following established procedures in previous studies to explore pupils’ language skills (Bowyer-Crane et al., 2017).

At follow-up, the original **sem** model to derive the secondary outcome measure – reading skills – was:

```
sem (READ_fu -> t3EWR_Total t3PC1_ReadFl t3PC2_ReadFl t3PC3_ReadFl t3PC1_ReadComp t3PC2_ReadComp
t3PC3_ReadComp), latent (READ_fu) standardized
```

The initial model fit was below accepted thresholds: CFI = 0.757; RMSEA= 0.335.

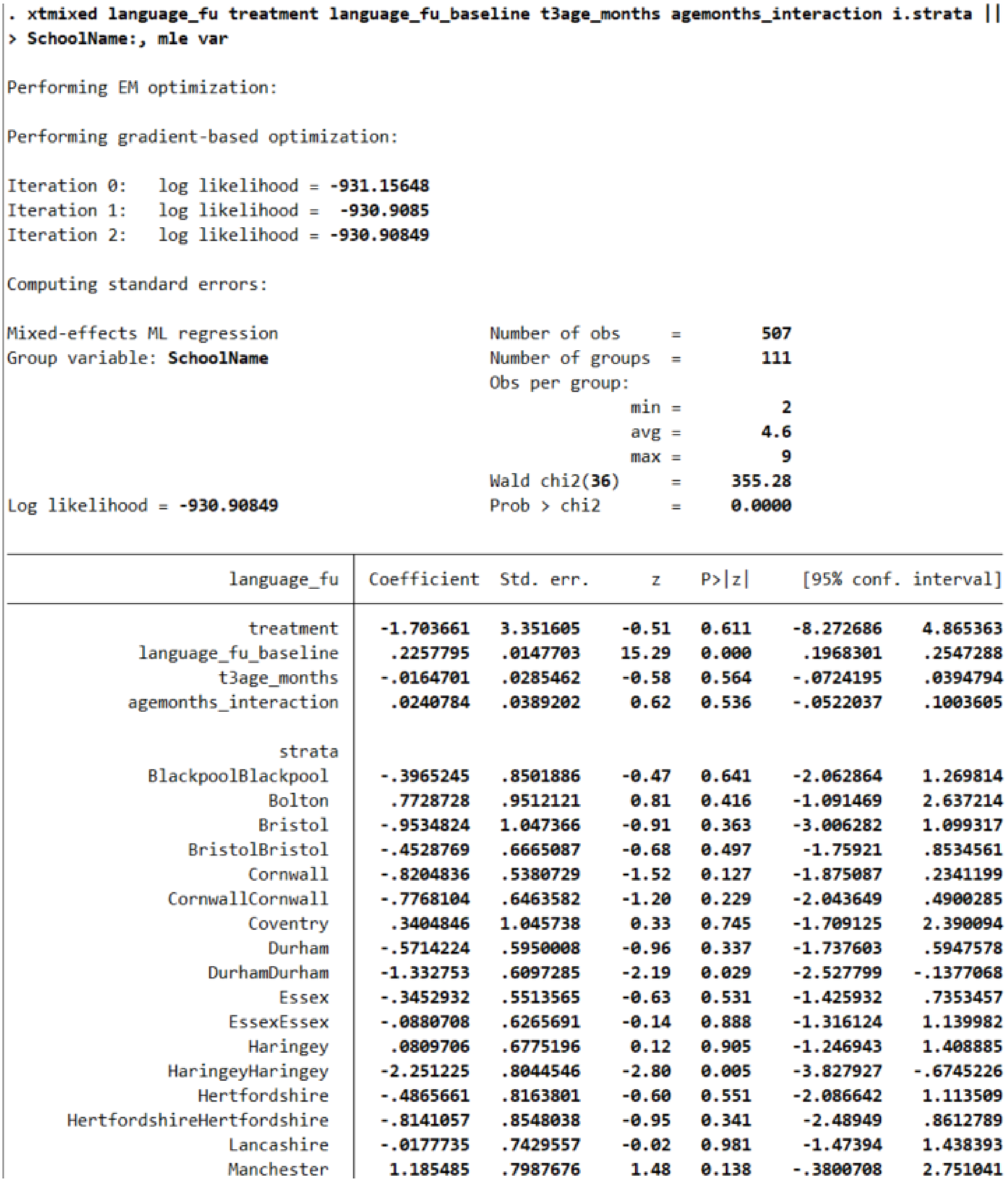
The model was therefore modified, adding two error covariances, in line with the theoretical assumption that both fluency indicators measure the same reading subskill and therefore their residuals can be expected to co-vary for reasons unrelated to the model. Similarly, both comprehension indicators are likely to co-vary due to a common factor that is not included in the model, but instead is specific to the pupils’ reading comprehension ability:

```
sem (READ_fu -> t3EWR_Total t3PC1_ReadFl t3PC2_ReadFl t3PC3_ReadFl t3PC1_ReadComp t3PC2_ReadComp
t3PC3_ReadComp), latent (READ_fu) standardized cov(e.t3PC2_ReadFl*e.t3PC3_ReadFl)
cov(e.t3PC2_ReadComp*e.t3PC3_ReadComp)
```

Model fit achieved after this one modification was outside the accepted cut-off points for both indices: CFI = 0.939; RMSEA = 0.182. However, further modifications to the model were deemed undesirable, as they could not be theoretically justified with sufficient confidence. Inspection of other model fit indices, such as the SRMR = 0.061 and TLI = 0.892 (with rules of thumb suggesting a cut-off value close to 0.08 for acceptable fit and 0.95, respectively) suggested a lack of robustness in the hypothesised model and therefore no further specifications were tested.

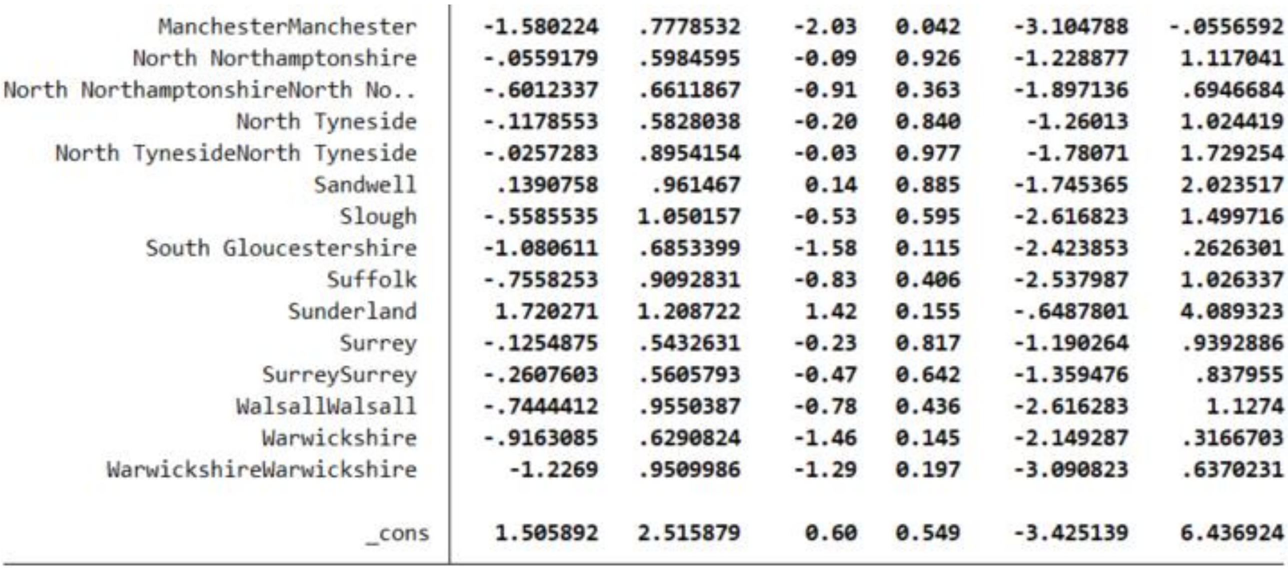
As a result, it was concluded that the three YARC tests combined did not create a stable latent reading variable using SEM and the three YARC tests were each analysed independently (for discussion see *Approach to SEM* section).

Appendix D: Sensitivity Check on Age at Follow-up

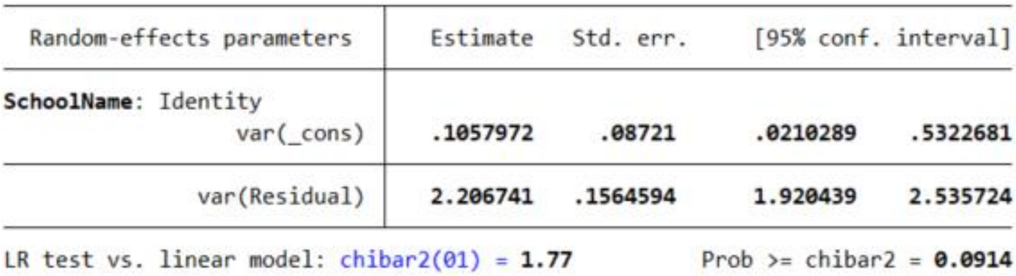


```
. xtmixed language_fu treatment language_fu_baseline t3age_months agemonths_interaction i.strata ||
> SchoolName:, mle var

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood = -931.15648
Iteration 1:   log likelihood =  -930.9085
Iteration 2:   log likelihood = -930.90849

Computing standard errors:

Mixed-effects ML regression                     Number of obs     =        507
Group variable: SchoolName                      Number of groups  =        111
                                                Obs per group:
                                                              min =          2
                                                              avg =        4.6
                                                              max =          9
                                                Wald chi2(36)     =     355.28
Log likelihood = -930.90849                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------------
                 language_fu | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-----------------------------+------------------------------------------------------
                   treatment |  -1.703661   3.351605    -0.51   0.611    -8.272686    4.865363
        language_fu_baseline |   .2257795   .0147703    15.29   0.000     .1968301    .2547288
                t3age_months |  -.0164701   .0285462    -0.58   0.564    -.0724195    .0394794
       agemonths_interaction |   .0240784   .0389202     0.62   0.536    -.0522037    .1003605
                             |
                      strata |
          BlackpoolBlackpool |  -.3965245   .8501886    -0.47   0.641    -2.062864    1.269814
                      Bolton |   .7728728   .9512121     0.81   0.416    -1.091469    2.637214
                     Bristol |  -.9534824   1.047366    -0.91   0.363    -3.006282    1.099317
              BristolBristol |  -.4528769   .6665087    -0.68   0.497     -1.75921    .8534561
                    Cornwall |  -.8204836   .5380729    -1.52   0.127    -1.875087    .2341199
            CornwallCornwall |  -.7768104   .6463582    -1.20   0.229    -2.043649    .4900285
                    Coventry |   .3404846   1.045738     0.33   0.745    -1.709125    2.390094
                      Durham |  -.5714224   .5950008    -0.96   0.337    -1.737603    .5947578
                DurhamDurham |  -1.332753   .6097285    -2.19   0.029    -2.527799   -.1377068
                       Essex |  -.3452932   .5513565    -0.63   0.531    -1.425932    .7353457
                  EssexEssex |  -.0880708   .6265691    -0.14   0.888    -1.316124    1.139982
                    Haringey |   .0809706   .6775196     0.12   0.905    -1.246943    1.408885
            HaringeyHaringey |  -2.251225   .8044546    -2.80   0.005    -3.827927   -.6745226
               Hertfordshire |  -.4865661   .8163801    -0.60   0.551    -2.086642    1.113509
  HertfordshireHertfordshire |  -.8141057   .8548038    -0.95   0.341     -2.48949    .8612789
                  Lancashire |  -.0177735   .7429557    -0.02   0.981     -1.47394    1.438393
                  Manchester |   1.185485   .7987676     1.48   0.138    -.3800708    2.751041
```

(results continued overleaf)



```
     ManchesterManchester |  -1.580224   .7778532    -2.03   0.042    -3.104788   -.0556592
     North Northamptonshire |  -.0559179   .5984595    -0.09   0.926    -1.228877    1.117041
North NorthamptonshireNorth No.. |  -.6012337   .6611867    -0.91   0.363    -1.897136    .6946684
            North Tyneside |  -.1178553   .5828038    -0.20   0.840     -1.26013    1.024419
  North TynesideNorth Tyneside |  -.0257283   .8954154    -0.03   0.977     -1.78071    1.729254
                  Sandwell |   .1390758    .961467     0.14   0.885    -1.745365    2.023517
                    Slough |  -.5585535   1.050157    -0.53   0.595    -2.616823    1.499716
       South Gloucestershire |  -1.080611   .6853399    -1.58   0.115    -2.423853    .2626301
                   Suffolk |  -.7558253   .9092831    -0.83   0.406    -2.537987    1.026337
                Sunderland |   1.720271   1.208722     1.42   0.155    -.6487801    4.089323
                    Surrey |  -.1254875   .5432631    -0.23   0.817    -1.190264    .9392886
              SurreySurrey |  -.2607603   .5605793    -0.47   0.642    -1.359476     .837955
            WalsallWalsall |  -.7444412   .9550387    -0.78   0.436    -2.616283      1.1274
              Warwickshire |  -.9163085   .6290824    -1.46   0.145    -2.149287    .3166703
    WarwickshireWarwickshire |    -1.2269   .9509986    -1.29   0.197    -3.090823    .6370231
                           |
                     _cons |   1.505892   2.515879     0.60   0.549    -3.425139    6.436924
------------------------------------------------------------------------------------------
```



```
------------------------------------------------------------------------------
  Random-effects parameters  |   Estimate   Std. err.     [95% conf. interval]
-----------------------------+------------------------------------------------
SchoolName: Identity         |
                  var(_cons) |   .1057972     .08721      .0210289    .5322681
-----------------------------+------------------------------------------------
               var(Residual) |   2.206741   .1564594      1.920439    2.535724
------------------------------------------------------------------------------
LR test vs. linear model: chibar2(01) = 1.77          Prob >= chibar2 = 0.0914
```

The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Education.

This document is available for download at https://educationendowmentfoundation.org.uk

Education
Endowment
Foundation

The Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
London
SW1P 4QP

https://educationendowmentfoundation.org.uk

@EducEndowFoundn

Facebook.com/EducEndowFoundn