Education Endowment Foundation

URLEY

Evaluation Report

February 2020

Hazel Wright, Dan Carr, Juliane Wiese, Lucy Stokes, Johnny Runge, Richard Dorsett, Jessica Heal, Jake Anders

THE BEHAVIOURAL INSIGHTS TEAM

The Education Endowment Foundation (EEF) is an independent grant-making charity dedicated to breaking the link between family income and educational achievement, ensuring that children from all backgrounds can fulfil their potential and make the most of their talents.

The EEF aims to raise the attainment of children facing disadvantage by:

- identifying promising educational innovations that address the needs of disadvantaged children in primary and secondary schools in England;
- evaluating these innovations to extend and secure the evidence on what works and can be made to work at scale; and
- encouraging schools, government, charities, and others to apply evidence and adopt innovations found to be effective.

The EEF was established in 2011 by the Sutton Trust as lead charity in partnership with Impetus (formerly Impetus Trust) and received a founding £125m grant from the Department for Education.
Together, the EEF and Sutton Trust are the government-designated What Works Centre for improving education outcomes for school-aged children.

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

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

0207 802 1653

jonathan.kay@eefoundation.org.uk

www.educationendowmentfoundation.org.uk

Contents

About the evaluator

Hazel Wright (The Behavioural Insights Team): Principal investigator, project management, and impact evaluation.

Dan Carr (The Behavioural Insights Team): Lead on impact evaluation design, statistical methods, and project management.

Juliane Wiese (The Behavioural Insights Team): Analysis and impact evaluation.

Lucy Stokes (The National Institute of Economic and Social Research): Analysis, impact evaluation, and advice on design and statistical methods.

Johnny Runge (The National Institute of Economic and Social Research): Implementation and process evaluation.

Richard Dorsett (The University of Westminster): Advice on design and analysis, impact evaluation.

Jessica Heal (The Behavioural Insights Team): Implementation and process evaluation lead.

Jake Anders (UCL): Advice on design and analysis.

Michael Sanders (formerly the Behavioural Insights Team) was the principal investigator on the evaluation until December 2018. Louise Jones (also formerly the Behavioural Insights Team) also provided support on implementation of the evaluation. Anitha George (formerly NIESR) contributed to the initial stages of the design for the IPE.

Executive summary

The project

The URLEY programme—Using Research Tools to Improve Language in the Early Years—trains teachers to improve children's language and social-behavioural outcomes in nursery and reception year (ages 3 to 5). Early years teachers take part in five-day-long professional development workshops in which they are introduced to a set of evidence-based language learning principles, taught how to use research tools (primarily the Environment Rating Scales, 'ERS') to assess their practice, and provided with strategies for refining practice. Mentors supported teachers to implement the approach in their schools using face-to-face and distance (skype/phone) sessions. A follow-up workshop in the third term was offered to review progress, consolidate learning, and plan next steps.

One hundred and twenty primary schools from the West Midlands, Liverpool, and Manchester participated in this efficacy trial from October 2016 to July 2018; 1,978 children were included in the evaluation. The programme was evaluated using a randomised controlled trial, testing the impact of the URLEY programme on children's language development over two years—compared to business as usual in control schools—using a composite language assessment. Children's social-behavioural development and the quality of practice in the participating settings were also assessed. The intervention was developed and delivered by a team from Oxford University, University College London (UCL) Institute of Education, and A+ Education. Interviews, case studies, and a survey were conducted to explore how the programme was implemented and to obtain feedback from participants.

Key conclusions
1. Children in schools receiving URLEY did not make additional progress in language development compared to children in control schools, as measured by a composite language score. This finding has a moderate to high security rating. The effect size is equivalent to one month's less progress than the control group, though is equivalent to zero months once imbalance on the numbers of FSM and EAL children in each arm is controlled for. The result was similar for pupils eligible for FSM.
2. The programme had a positive impact on quality of provision (as measured by Environment Rating Scales), with effect sizes in the range of 0.5–0.7. This suggests that quality of practice improved (for example, the quality of language-supporting adult-child interactions) but not at a sufficient level to translate to improved language outcomes for children. It may be that impacts on pupil outcomes would only be observed in the longer term, or with even larger improvements to practice.
3. Many children were not taught in reception by a teacher who had received the full training (partly due to substantial teacher turnover in the schools); it was not possible to assess the extent and impact of this in the evaluation. Additional induction training was provided where possible, but this is nonetheless likely to have reduced the potential impact of the URLEY programme.
4. Teachers were overwhelmingly positive about the URLEY programme: 91% of responding teachers felt the intervention had a positive impact on the quality of provision and highlighted the mentoring as especially valuable. Many teachers felt the programme was most beneficial to a targeted subset of reluctant communicators, as opposed to whole-class improvements.
5. The URLEY programme required significant time investment and it was found that cascading the intervention to staff who did not attend the training was challenging. Condensed training and a more structured approach with milestones, goals, and senior leadership team (SLT) support may have helped teachers to prioritise the programme.

EEF security rating

The primary finding has a moderate to high security rating. This was an efficacy trial, which tested whether the intervention worked under developer-led conditions in a large number of schools. It was a well-designed randomised controlled trial, however five schools withdrew from the trial after randomisation and 22% of pupils had missing data, which reduces the security rating. The main results were substantively unchanged by a range of sensitivity analyses to account for missing data. Although the pre-specified headline result shows a negative direction, there was essentially no difference in progress between treatment and control groups once imbalance on the numbers of FSM and EAL children in each arm was controlled for.

Additional findings

The evaluation did not find a positive effect on the composite language outcome or on its subcomponents—the British Picture Vocabulary Scale (BPVS), the Renfrew Action Picture Test (RAPT) Information, RAPT Grammar, and the Clinical Evaluation of Language Fundamentals (CELF). There was also no meaningful impact on social-behavioural development as measured by the Adaptive Social Behaviour Inventory (ASBI), though the results showed substantial ceiling and floor effects to the ASBI scales, which mean it would have been difficult for any intervention to effect a change on these outcomes.

The evaluation found a positive impact of the programme on each of the four reported measures of practice quality as measured by the Environment Rating Scales (ERS). The largest effect size (0.7, CI 0.41, 0.96) was identified for a composite of items relating to oral language development focusing primarily on language-supporting adult-child interactions. Smaller effect sizes (0.5–0.6) were identified on overall quality as measured by the ECERS-3, SSTEW, and ECERS-E literacy subscale indicating that the intervention had had a knock-on impact on other aspects of provision. This supports the programme's theory of change but poses a question of why this did not translate into improved language outcomes. One possible explanation is that quality ratings post-intervention were still in the minimal-to-adequate range and that child impacts might be seen only once settings reach a higher quality level, or have embedded the changes over a longer time. Or, it may be that individual children did not receive enough of the improved language interactions to make a difference to their development, for example, because of the relatively low adult-child ratios in schools, or because they were in classes where teachers had not received the full training.

Despite the time commitment required, the process evaluation revealed a high level of teacher enthusiasm for the programme: 88% of respondents reported large positive impacts for themselves as teachers (for example, in their practice, professional vision, and child development knowledge) and perceived benefits for children. The Language Principles were used more frequently and reported to improve practice; the ERS were used less often, though may have helped teachers become more mindful of the components of classroom quality and contributed to the higher ERS scores.

Cost

Delivering the URLEY intervention costs £3,885 per school, or £51.80 per pupil per year averaged over three years. The majority of the costs are realised in the first year.

Impact

Table 1: Summary of impact on primary outcome

Outcome/ Group	Effect size (95% confidence Interval)	Estimated months' progress	EEF security rating	No. of pupils	P value	EEF cost rating
Composite language	-0.08 (-0.19, 0.03)	-1	🔒🔒🔒🔓🔓	1,978	0.15	£ £ £ £ £

Introduction

Background evidence

The Using Research Tools to Improve Language in the Early Years (URLEY) intervention is an evidence-based professional development programme for nursery and reception teachers. It is designed to improve participants' knowledge of how children learn and develop oral language skills, and how to support that learning through evidence-based practice, including use of research tools such as the Environment Rating Scales (ERS) to evaluate and refine pedagogy and practice. Its primary aim is to improve children's language skills. This report presents findings from the evaluation of the URLEY intervention.

The ERS provide a framework through which the early learning environment can be understood and assessed, focusing on child-centred pedagogy. The longitudinal Effective Provision of Pre-school, Primary and Secondary Education (EPPSE) study found that observed ratings of quality from the ECERS, controlling for other contributing factors, were associated with improved attainment in maths and English as well as better social outcomes, through to the end of Key Stage 2 (Sylva et al., 2008). Even at age 16, attending a higher quality pre-school (as measured by ECERS) was associated with better performance at GCSE as well as improved self-regulation and pro-social behaviour (Sylva et al., 2014). Analysis of a sub-study of early years settings attended by children in the Millennium Cohort Study also showed a positive association between higher ECERS scores and improved language outcomes (Hopkin, Stokes and Wilkinson, 2010). One limitation of this literature, however, is that these studies show evidence for correlation rather than causation. Moreover, these studies explore the relationship between observations made using the ERS and children's outcomes. In contrast, the URLEY intervention supports nursery and reception teachers to understand the principles behind the ERS and to use a range of tools and resources to evaluate the extent to which their practice is evidence based in order to make refinements. The current evaluation therefore aims to determine if training nursery and reception staff to better understand child development and pedagogy, and to use evidence-based frameworks to evaluate and improve their practice, will lead to improved attainment in children's language and social development.

There is robust evidence demonstrating the efficacy of interventions targeted during early childhood in improving both cognitive and non-cognitive abilities. In the cognitive space, early intervention programmes have been linked to improved measures of school readiness, academic attainment, and school progression (Anderson et al., 2003). Participation in such programmes has also been associated with positive outcomes beyond school years, including lower rates of interaction with the criminal justice system, reduced likelihood of teenage pregnancy, reduced welfare dependency, and lower prevalence of risk factors for cardiovascular and metabolic diseases (Campbell et al., 2014; Currie, 2001; Gorey, 2001). Additionally, early interventions such as reducing pupil-teacher ratios can provide a large return on investment, particularly compared to those occurring in later years (Heckman and Masterov, 2007; Heckman, 2012).

In the preschool context, several studies have investigated the impact of training teaching staff in new practices or curricula. U.S. studies have shown that training and supporting preschool teachers in classroom management strategies can reduce internalising and externalising behaviour problems in children and reduce signs of teacher-reported social withdrawal (Raver et al., 2009). Similarly, training preschool teachers in curricula designed to increase child social-emotional competence has been linked to improved child emotional knowledge skills and parent and teacher-reported measures of social competence (Domitrovich, Cortes and Greenberg, 2007; Webster-Stratton, Reid and Stoolmiller, 2008), as well as improved self-regulation (Webster-Stratton, Reid and Stoolmiller, 2008).

Studies have also demonstrated the effectiveness of interventions in the early years for language development (Springate et al., 2008). The Nuffield Early Language Intervention, comprising staff training (principally for teaching assistants), lesson plans, and materials, was found to have a positive impact on children's language skills (Sibieta et al., 2016). This intervention targeted children with poor spoken language skills; however, there is also evidence that non-targeted programmes can be effective for improving language outcomes (Springate et al., 2008).

Of particular relevance to the present study, there is evidence from the U.S. that promoting the capacity of teachers to use research-informed tools to improve their practice can advance both child social-emotional and cognitive development. The 'Research Based, Developmentally Informed' (REDI) clustered randomised controlled trial equipped Head Start preschool teachers with a portfolio of lesson plans and enrichment activities. Teachers also received

substantial training in all materials and guides provided, and were supported by weekly meetings with a mentor. One year after the trial began, children who received the intervention had progressed significantly further in measures of vocabulary, emergent literacy, and socio-emotional development (Bierman et al., 2008). A follow-up study tracking the children into kindergarten (one-year after the first study concluded) found evidence for a sustained intervention effect across most measures of socio-emotional development, but only detected a continuing significant effect in one measure of language development, phonemic decoding (Bierman et al., 2014).

Currently, the URLEY intervention is being used in schools in two Strategic School Improvement Fund (SSIF) projects, funded by the DfE, and is also offered by A+ Education Ltd to settings and schools. For the purpose of this trial, it has been delivered to primary schools with nursery and reception classes across England. Since the end of the intervention period, the delivery team estimate that 460 additional practitioners have received support.

Intervention

The URLEY programme is designed to support teachers in using pedagogical assessment tools to evaluate and improve the quality of language-supporting practice. The description below follows the Template for Intervention Description and Replication (TIDiER) checklist.

1. Brief Name: The Using Research Tools to Improve Language in the Early Years programme (URLEY).

2. Why (rationale/theory): The URLEY intervention aims to 'support participating nursery and reception class teachers in improving children's language skills by using research-validated tools to:

- improve their knowledge of how children learn and develop oral language, and how to support that learning through evidence-based practice;
- use language and pedagogical assessment tools to "tune in" to children and practice, acting on evidence gathered to improve the effectiveness of teaching;
- involve their wider classroom teams in developing the quality of language-supporting practice via pedagogical leadership;
- develop an effective self-evaluation and improvement cycle, and improve confidence to articulate shared pedagogy to others; and
- build capacity for sustained improvement beyond the end of the programme.'

The language domains of focus are: social communication, vocabulary, grammar, and narrative skills. The URLEY programme is underpinned by research relating to language acquisition and effective language-supporting pedagogy. These are summarised within nine language-learning principles (based on a wide-ranging review of the current literature):

- be a magnet for communication;
- create irresistible and meaningful contexts for communication;
- support language at home;
- be a language radiator;
- create a culture of adult-child conversation;
- support children to communicate and listen to each other;
- create repeated opportunities for children to bump into and use new words;
- offer children clear information about word meanings; and
- provide sensitive and meaningful feedback on children's language.

The programme also draws heavily on theory and research relating to adult learning and the effective characteristics of Continuous Professional Development, including research by Cordingley et al. (2015), Timperley et al. (2007), and Stoll et al. (2012). Evidence-based features include:

1. having a specific and articulated objective—starting with the end in mind in relation to the specific domains of oral language development to be promoted;
2. an explicit focus on practice and on linking knowledge and theory to practice;
3. an intensity and duration matched to the content being conveyed—evidence suggests that two terms or more are required when the objective is to achieve significant organisational and cultural change;

4. support for practitioners in conducting child assessments and interpreting their results as a tool for ongoing monitoring of the effects of professional development;
5. approaches that are appropriate for the organisational contexts of participating schools and settings and are aligned with standards for practice;
6. active rather than passive learning approaches;
7. collective participation of teachers from the same classrooms or schools;
8. access to expert knowledge; and
9. high quality content.

The programme had a secondary focus on supporting children's personal, social, and emotional development alongside language development.

3. What (materials): Resources and tools to support self-evaluation and improvement included:

- the Environment and Quality Rating Scales—research-validated observational rating scales known to predict aspects of children's development; three scales were used: the ECERS-3, the ECERS-E, and the SSTEW (described in further detail below);
- videos of effective pedagogical practice watched during training workshops;
- research readings and other evidence-based resources to further deepen knowledge and understanding of oral language and pedagogy and provide evidence-based strategies for implementation;
- the Teacher Rating of Oral Language and Literacy (TROLL, Dickinson, McCabe and Sprague, 2003), a short oral language screening tool that tracks children's language and literacy development; and
- online resources—including the DVD clips used during the training, programme implementation materials, and a Pinterest board.

4. What (procedures): The model tested in this programme comprised six days of Continuing Professional Development (CPD) on oral language development for nursery and reception class teachers, provided over the course of 11 months (February to October 2018). A further three days (on average) of in-class mentoring and coaching was provided to each participating school in order to support implementation. During training, teachers were introduced to a set of evidence-based language learning principles, taught how to use research tools (primarily the ERS) to evaluate the extent to which these principles were currently being implemented in their classroom, and provided with tools and strategies for refining practice based on their evidence.

The core components of the intervention were:

- A set of underpinning Language Learning Principles summarised the research evidence on how children learn language and how their development can best be supported within the classroom.
- Research readings and other evidence-based resources were used to further deepen knowledge and understanding of both language and pedagogy, and provide evidence-based strategies for implementation.
- Research tools were used to support self-evaluation and improvement. The primary tools used were the Environment and Quality Rating Scales—research-validated observational rating scales known to predict aspects of children's development, with higher scores linked to improved maths and English attainment (Sylva et al., 2008). Three scales were used:

    1. The ECERS-3—the Early Childhood Environment Rating Scale, Third Edition—allows teachers to assess the quality of the classroom environment. The scale comprises 35 items organised into six subscales examining space and furnishings, personal care routines, language and literacy, learning activities, interaction, and programme structure. Scores are based on a three-hour classroom observation, with items in each subscale scored between one and seven (higher scores being more positive). Subscale and average scores are calculated by averaging all of the item scores.
    2. The ECERS-E—a version of the ECERS—was designed as an extension to the ECERs-3 to provide additional items to dimensions of quality for four other aspects of educational provision. These are: literacy, mathematics, science and environment, and diversity. It is scored in the same way, with higher scores indicating greater quality.
    3. The SSTEW—the Sustained Shared Thinking and Emotional Wellbeing scale—measures the quality of practices in the classroom that support the development of strong relationships, emotional wellbeing, shared thinking, self-regulation, and communication. It focuses on the pedagogy within the setting and the adult's role in supporting learning and development, as well as the quality of interactions between children. Items are assessed on a seven-point scale and are divided into five subscales and 14 items:

building trust, confidence and independence, social and emotional wellbeing, supporting and extending language and communication, supporting learning and critical thinking, and assessing learning and language. Higher scores are indicative of higher quality.

4. A particular focus was placed on the items addressing support for children's oral language skills and socio-emotional development. During training, teachers watched videos of effective practice and were supported to use the language principles and ERS to 'tune in' to language-supporting practice. Back in their classrooms, they were encouraged to use these same tools to observe and evaluate their use of evidence-based practices, and to identify potential improvements.

- Support was given for assessing children's language skills. Teachers were introduced to the Teacher Rating of Oral Language and Literacy (TROLL, Dickinson, McCabe and Sprague, 2003), a short screening tool, and asked to complete the oral language items for their class before the start of the CPD programme. TROLL offers teachers the opportunity to systematically track their students' language abilities and interests. The full tool consists of 25 questions covering language use, reading, and emergent writing. Its completion requires only five to ten minutes per child and can help provide a broader snapshot of the overall strengths and needs of the classroom (Dickinson, et al., 2003). The oral language assessments were then used throughout the CPD to inform improvement work and support for specific children. The TROLL provided a common tool for teachers to use throughout the programme. Teachers were also encouraged to use existing assessments (for example, their EYFS records) to inform their improvement work and support for oral language.
- Time for reflection, discussion, and sharing of practice during training days was facilitated with a focus on implementation of new strategies within the classroom.
- Time for implementation (including involving other staff) was also provided between workshops, with specific classroom-based activities and in-school mentoring to scaffold implementation.
- Finally, support was provided in developing an action-plan and improvement cycle, and in planning for language development.

5. Who (implementers): At least one nursery and one reception class teacher from each school were expected to attend the training days, with additional places available (and encouraged) for schools with more than one-form entry. In all, 148 participants were listed on the study database within 60 intervention schools—a mean of 2.5 teachers per school. Schools had on average 3.4 nursery and reception class teachers. Thus, approximately 74% of all possible teachers/classes participated. These teachers were expected to cascade the training to other staff within their classrooms and involve them in implementation.

Training was provided in six cohorts of up to 25 teachers. Since the programme ran over two academic years, significant turnover was seen in staffing (only 56% of listed participants were constant throughout the intervention period). A one-day induction training session was offered for new staff in September 2018. Teacher turnover is discussed in the IPE section of the report.

6. How (mode of delivery): Nursery and reception class teachers attended five full days of training over five months (approximately one day per month between February and June 2018) with a follow-up day in October 2018:

- Day 1: Tuning in (an introduction to the URLEY approach, child assessment tools, and Environment Rating Scales).
- Day 2: Language through play.
- Day 3: Supporting children's language skills.
- Day 4: Using books.
- Day 5: Supporting language through investigation, problem-solving, and exploration.
- Day 6: Follow-up day—review, refresher, planning for sustaining improvement work.

Schools also received mentoring and coaching to help them apply the programme in practice. Support was offered through a mix of face-to-face and distance (skype/phone) sessions. Each school individually received on average 17.5 hours of mentoring, through a combination of face to face and online support. Mentors were qualified teachers who had experience in relevant leadership or mentoring roles (for example, as a headteacher or deputy headteacher in a school, in a local authority, or other school improvement role). They received specific training to offer mentoring as part of the URLEY programme, including specialist input on oral language development, on the URLEY approach and materials, and on the research tools used in the intervention. Mentors attended regular meetings and group supervision sessions, and also received individual phone supervision throughout the period of the intervention.

7. Where (setting): CPD training days took place in a location outside the classroom, with participants attending the cohort which was most convenient for them geographically. Teachers from the same school attended the same cohort. Trial schools were based in Liverpool, Manchester, and the West Midlands.

8. When and how much (dosage): Nursery and reception class teachers attended six days of training in total between February and October 2018. For mentoring and coaching, all schools received an initial face-to-face visit of approximately one and a half hours. Following the initial visits, schools received an average of just over five face-to-face visits each. Visit times ranged from one hour to a full day, with a mean visit time of three hours. Support was provided at the school level, and mentoring divided between participating classes. Distance mentoring contact was shorter—less than an hour on average—with schools receiving an average of 1.4 support sessions each over the intervention period. On average, schools each received 17.5 hours of mentoring support.

9. Tailoring: Tailoring took place throughout delivery involving mentors providing support in translating new learning into practice, engaging other staff (such as teaching assistants) in the change process, and providing feedback through observation using the ERS. Such tailoring was seen as an acceptable part of the intervention; adaptations are discussed as part of the findings from the implementation and process evaluation.

10. Modifications: Following high levels of teacher turnover within intervention schools, a series of briefing sessions were offered in September 2018 for teachers new to the programme.

11. How well (planned): Both workshop trainers were involved in the development of the intervention and so had a deep knowledge of the material. Detailed trainer notes were provided to support fidelity, including slide notes, activity guides, and guidance to support discussion and analysis of DVDs and readings. Data was kept on any variations from the planned schedule.
A handbook was also provided for mentors, and regular group supervisory meetings were held (approximately one face-to-face session per month, with shorter skype sessions where needed). Mentors were also provided with individual supervisory phone-calls conducted by a senior member of the research team.

Detailed fidelity information was gathered on the mentor visits, including visit focus (set by staff, collaborative decision, suggested by mentor), the techniques used (using a common framework), and visit timings and members of staff involved. Data was also gathered on postponements and cancellations of mentor visits by schools.

Evaluation objectives

The aim of the evaluation was to assess the impact of the URLEY intervention on language and social-behavioural outcomes for nursery and reception year pupils aged three to five.

The primary research question was:

- Does the URLEY intervention improve language development over the nursery and reception years of schooling as assessed by a battery of language measures, namely the British Picture Vocabulary Scale (BPVS), Renfrew Action Picture Test (RAPT), and Clinical Evaluation of Language Fundamentals (CELF) Preschool 2 UK—Sentence Structure subset?

The impact evaluation also aimed to address the following secondary research questions:

- What is the impact of the programme on children's social-behavioural outcomes as measured by the Adaptive Social Behaviour Inventory (ASBI) score?
- Does the programme impact on the quality of provision for language and social-behavioural development as measured by a composite of items from the ECERS-3, ECERS-E, and SSTEW—collectively the 'Environment Rating Scales' (ERS)?
- What is the impact of the programme on children's language and social outcomes for pupils who move to a reception class with a teacher who did not directly receive ERS training and mentoring?
- What is the impact of the programme on children's language and social outcomes for pupils eligible for free school meals?
- What is the impact of the programme on children's language and social outcomes for pupils for whom English is an Additional Language?
- What is the impact of the programme on children's language and social outcomes for pupils who have language difficulties at the start of the trial?
- Does the programme show differences in impact on children's language and social outcomes according to children's gender?

The implementation and process evaluation (IPE) aimed to take a triadic approach through the analysis of the implementation of the intervention, delivery of the intervention, and perceived impact of the intervention. The IPE addressed the following main research questions specific to the intervention:

- To what extent did the intervention training and mentoring support teachers to develop a research-led evaluative mindset and use practical strategies to implement the ERS items within their classroom?
- To what extent did teachers engage with the intervention as a tool to catalyse improvement in language and social-behavioural outcomes for children?

These questions were underpinned by a series of sub-questions relating to both intervention delivery and engagement with the intervention, as set out in the evaluation protocol.
The evaluation protocol can be found here. The statistical analysis plan can be found here.

Ethics and trial registration

Ethical review was undertaken by the University of Oxford with the project receiving ethical approval from the Departmental Research Ethics Committee in April 2016. In addition, NIESR adheres to the Ethics Guidelines of the Social Research Association.

As the intervention was delivered within school hours, consent from the school was considered sufficient with regard to consent for the intervention, and as randomisation took place at the school level, the decision to enter into randomisation could also be made by the school.

The requirements for schools of participating in the study were set out in a Memorandum of Understanding (MoU) for the evaluation, which all participating schools were required to sign. A copy of the MoU is provided in Appendix C. Information about the evaluation was also provided in an information sheet for schools; in addition, recruitment events offered the chance for schools and teachers to find out more about the project.

Given the young age of the pupils involved in the study, and the one-to-one language assessments that would be carried out, an opt-in consent process was used, with participants' parents (or legal guardians) making an informed decision regarding whether they consented to their child's participation in the assessments and data-sharing. An information leaflet was sent to parents along with the opt-in consent form (Appendix C) providing information on the aims of the research and the use of data so as to allow parents to make an informed decision.

Consent letters were returned by parents to their child's school, with schools then returning these to BIT. While some schools stated that they had posted consent forms to BIT, in two cases these did not arrive. All schools provided written confirmation that they had posted the forms. Furthermore, schools were previously asked—as part of the process of collecting pupil data from schools—to confirm that they had only included details for pupils for whom consent had been obtained. In June 2017, the evaluation and delivery team sought the advice of the University of Oxford Departmental Research Ethics Committee on this matter. The request for this updated approval was made in July 2017 and granted in October 2017. As a result, all schools where consent forms were missing were re-contacted and asked to re-supply consent forms. Ultimately, for the two schools that did not provide any consent forms, all child data was removed from the evaluation. A further 114 pupils from other schools were removed from the evaluation where consent had not been received.

This trial has been registered at: https://doi.org/10.1186/ISRCTN18055918

Data protection

As noted above, schools were provided with an information sheet setting out the details of the evaluation prior to deciding whether to take part, with the requirements of participation detailed in the MoU.

Given the young age of the pupils involved in the study, a parental opt-in consent process was used, with participants' parents (or legal guardians) making an informed decision regarding whether they consented to their child's participation in the assessments and data sharing. Alongside an opt-in consent form, parents were provided with the right to be informed in the form of an information leaflet explaining the trial, describing the team's reasons for seeking access to their child's data and an explanation of how this data would be used. Specifically, this information highlighted their right to object and their right to erase their child's dataset from the evaluation. Parents were informed that they could choose

not to take part and stop at any time, that they could ask for data to be removed from the study at any time, and that data would be linked to the National Pupil Database (held by the Department for Education) and shared with the evaluation team, the research team, the Department for Education, the EEF, the EEF's data contractor, FFT Education, and (in an anonymised form) to the U.K. Data Archive. The legal basis for data collection and processing by the evaluation team prior to and under GDPR was fully informed opt-in consent.

Relevant consent documents can be found in Appendix C.

Project team

Development and delivery team

Sandra Mathers (University of Oxford): Senior researcher and study principal investigator.

Iram Siraj (University College London Institute of Education): Professor of Child Development and Education, and study co-investigator.

Clare Williams (A+Education): Director and study co-investigator.

Maria Evangelou (University of Oxford): Associate professor and study co-investigator.

The evaluation team

Hazel Wright (The Behavioural Insights Team): Principal investigator, project management, and impact evaluation.

Dan Carr (The Behavioural Insights Team): Lead on impact evaluation design, statistical methods, and project management.

Juliane Wiese (The Behavioural Insights Team): Analysis and impact evaluation.

Lucy Stokes (The National Institute of Economic and Social Research): Analysis, impact evaluation, and advice on design and statistical methods.

Johnny Runge (The National Institute of Economic and Social Research): Implementation and process evaluation.

Richard Dorsett (The University of Westminster): Advice on design and analysis, impact evaluation.

Jessica Heal (The Behavioural Insights Team): Implementation and process evaluation lead.

Jake Anders (UCL): Advice on design and analysis.

Michael Sanders (formerly the Behavioural Insights Team) was the principal investigator on the evaluation until December 2018. Louise Jones (also formerly the Behavioural Insights Team) also provided support on implementation of the evaluation. Anitha George (formerly NIESR) contributed to the initial stages of the design for the IPE.

Methods

Trial design

The trial utilised a two-arm cluster randomised controlled trial comparing the URLEY programme to a control condition with randomisation at school level and a post-test of language development across several key domains (collapsed into a single composite score for the primary outcome measure). See **Table 3** for a summary.

Table 3: Summary of the URLEY programme.

Trial type and number of arms		Two-arm, cluster randomised.
Unit of randomisation		School.
Stratification variable(s) (if applicable)		Geographic area: West Midlands, Manchester, and Liverpool. Proportion of FSM pupils: above or below mean in 2015/2016.
Primary outcome	variable	Language skills.
	measure (instrument, scale)	Composite language skill score, based on: ● British Picture Vocabulary Scale; ● Renfrew Action Picture Test; and ● Clinical Evaluation of Language Fundamentals (CELF) Preschool 2 UK – Sentence Structure.
Secondary outcome(s)	variable(s)	Variables comprised: ● each of the language measures used in the construction of the composite score; ● social-behavioural development; and ● quality of provision for language and social development.
	measure(s) (instrument, scale)	Individual language skill measures: ● British Picture Vocabulary Scale; ● Renfrew Action Picture Test (both information and grammar scores); and ● CELF Preschool 2 UK – Sentence Structure. Social-behavioural development: ● Adaptive Social Behaviour Inventory. Provision quality: ● composite environment rating scale score based on items from ECERS-3, ECERS-E, and SSTEW.

As the programme involves training teachers to improve practice in their classes, the choice was between randomising at school or class level. Randomisation at class level would have entailed substantial risk of cross-contamination, especially as part of the programme involves teachers sharing practice with other staff. Furthermore, as the trial aimed to follow children from nursery into their reception year, it would not have been practical to ask schools to keep class

groups the same when moving from nursery to reception as school's choices around class allocation are necessarily driven by many other factors.

The trial aimed to recruit 120 primary schools (ultimately 122 were recruited) with nursery and reception classes. The target was overshot as there was a lull in recruiting schools towards the randomisation deadline, and to ensure the target number was reached more schools than required were approached. This was necessary as past trials in the education space have shown that schools typically withdraw after volunteering to participate or fail to provide the pupil data necessary by required dates, ending their involvement in the trial.

Randomisation of schools occurred in two batches (first and second wave). Batch randomisation was not specified in the trial protocol, but became necessary as the majority of schools recruited had submitted all required information in advance of the original randomisation date and were eager to know their trial arm allocation in order to arrange cover for staff needing to attend training sessions if assigned to the treatment arm. Delaying randomisation for these schools risked several withdrawing from the study. Consequently, a decision was made in conjunction with the EEF to randomise those ready by the original randomisation date in one batch, and to randomise all those not ready at that date in a later batch. The need to randomise in batches meant the stratification variables detailed in the trial protocol were altered to account for the smaller number of schools being randomised in each instance (to avoid strata with only a single school). Rather than stratifying on school-level FSM share (split across the median sample proportion), local authority, and Key Stage 1 (KS1) English reading attainment (split across the median sample result), schools were stratified with respect to the proportion of FSM students (above or below the median sample proportion) and school location (West Midlands, Manchester, or Liverpool).

Schools in the control group were expected to continue with 'business as usual' and were offered the opportunity to take part in the programme following the completion of the study (August 2018), or a payment of £1,000, whichever they preferred. The £1,000 was intended to cover the majority of the costs for delivery of the intervention to two teachers per school.

Participant selection

The project delivery team was responsible for recruiting schools into the trial. The trial was conducted across schools in the Liverpool, Manchester, and West Midlands areas. These areas were chosen as they possess above average proportions of students from disadvantaged backgrounds (as measured by the proportion of neighbourhoods in the top 20% of areas in the Indices of Multiple Deprivation) and/or below average results at age five for communication and language (30% worst performing authorities for the proportion of children reaching expected level of development for communication and language in their EYFS profile).

The eligibility criteria for schools to participate were:

- Participating schools were to be located in one of the study areas, and be a one- or two-form entry state primary school with a nursery class. Three- or four-form entry schools were only accepted where they agreed to channel nursery children who had completed a baseline assessment into a reception class led by a participating teacher (defined as one nominated during the EOI process). This applied to both control and treatment schools. That is, in control group schools nursery children moved to a reception class led by a teacher who would have received the intervention had they been assigned to the treatment condition.
- One nursery and one to two reception teachers (with three reception teachers encouraged where the school was three-form or more entry) were to have agreed to attend the ERS training and engage with mentoring if allocated to the treatment group.
- Schools should not have previously accessed training by A+ Education—which is substantially similar to that being provided via the current intervention—received substantial support from their local authority using rating scales such as the Environment Rating Scales (ECERS and others), or used such tools themselves on a regular basis.
- If allocated to the control group, schools were required to continue with 'business as usual' for the duration of the trial (that is, not procure similar training that they otherwise would not have done).
- Schools should have signed a completed Memorandum of Understanding.
- Schools should have consented to participate in the study—including the collection of outcome measures in summer 2018—regardless of which trial arm they are assigned to.

- Schools should have agreed to obtain opt-in consent from the parents of children involved in the study, and to provide both school- and pupil-level data.
- Schools should have agreed to allow time for each assessment phase and liaise with the evaluation team to find appropriate dates and times for assessments to take place.
- Schools should have agreed that teachers in both trial arms complete a survey at the end of the trial period and attend an interview with evaluation staff if requested.

Seven hundred and thirty-two schools were approached to participate in the trial directly, with indirect approaches made to a wider pool of schools through Local Authorities; 152 schools responded with Expressions of Interest, of which 140 were eligible. Ultimately, 122 schools were recruited. Of schools that did not commit to the trial, reasons for not participating included (but were not limited to) staffing difficulties, lack of capacity for the evaluation, bad timing, and business with other programmes.

Outcome measures

Primary outcome

The primary outcome measure was a composite language skill score. This measure drew on the results of three language assessments, all of which were assessed on a one-to-one basis:

- **British Picture Vocabulary Scale (BPVS)**: This assessed a child's receptive vocabulary, or words that a person comprehends and responds to but cannot necessarily produce (Burger and Chong, 2011). For each question, the test administrator says a word and the child responds by selecting a picture from four options that best illustrate the word's meaning. The test typically takes ten minutes to administer.
- **Renfrew Action Picture Test (RAPT)**: In this test, the child is asked to describe the actions shown in a set of pictures. Two scores are recorded, one for the level of information they provide (for example nouns and verbs) and one for the grammar they use (such as use of tenses). Both information and grammar scores are incorporated into the composite language measure. We conducted moderation analysis on the RAPT which, compared to other tests, has more room for subjectivity when marking. This analysis, detailed in Appendix D, shows a strong correlation between Research Assistant (RA)-marked outcomes and moderated outcomes, suggesting that the scoring of outcome data was done to a reasonably high standard. The test typically takes five minutes to administer.
- **Clinical Evaluation of Language Fundamentals (CELF) Preschool 2 UK – Sentence Structure**: This subtest provides information about how a child understands spoken language. This is achieved by asking the child to interpret spoken sentences of increasing length and complexity by pointing to the picture that illustrates a given sentence. The test typically takes five minutes to administer.

These three assessments were chosen following extensive discussion with the project delivery team, taking into account the most appropriate measures for the aspects of language development targeted by the intervention, the age of children in the trial, and practical considerations, such as time needed to administer. Collectively the language assessments took between 15 and 20 minutes to administer.

Each assessment captured a different dimension of language development, covering vocabulary, comprehension, and expressive language. It was decided that the primary outcome should be a composite score based on these different measures, as a priori the intervention could potentially affect each of these dimensions. Although some of the component measures were different, they captured various elements of emerging literacy skills and are therefore appropriately combined. This approach was broadly consistent with that adopted in the evaluation of the Nuffield Early Language Intervention (Sibieta et al., 2016).

To arrive at a composite language skill score we standardised each of the components (including both RAPT scores) to have a mean of zero and standard deviation of one. These were added together to create a composite measure and re-standardised. As such, the four[1] language measure scores were equally weighted in the composite language skill score. In addition, we also explored factor analysis as a sensitivity analysis with each element of the composite score.

[1] As this includes the two separate RAPT measures, along with the scores from the BPVS and the CELF Sentence Structure sub-test.

All three assessments were administered at both pre- and post-test. All were administered on a one-to-one basis by research assistants (RAs) with an academic background in speech therapy or psychology. Recruited by BIT, the RAs were trained by an experienced language development psychologist in how to administer and score the language assessments prior to visiting schools. RAs were trained to cease administering the assessment if children showed irritation or no longer wished to continue. RAs were blind to the trial arm allocation of schools they visited to conduct assessments.

Assessments were conducted at two intervals during the course of the trial:

- **pre-test**—this was conducted between 5 October and 12 December 2016; the great majority occurred prior to schools being informed of their trial arm assignment, but in the case of five schools, some follow-up visits occurred one to ten days after trial-arm assignment was communicated to schools in order to include children who were absent when RAs first visited; and
- **post-test**—this was conducted during June and July 2018.

Secondary outcomes

Individual scores for each of the above language measures were also reported as secondary outcome measures.

An additional secondary outcome was social-behavioural development as measured by the Adaptive Social Behavior Inventory (ASBI). The ASBI is a questionnaire that was completed by class teachers for each child, at the same time-point as primary outcome measures were collected (at both pre- and post-test). We report scores for the three subscales, Express (13 items), Disrupt (7 items), and Comply (10 items), the Prosocial score (effectively the sum of the Express and Comply scores), as well as the total score (minimum score = 30, maximum = 90). As trial arm allocation was only revealed after pre-test outcome measures were collected, the pre-test ASBI scores were blind to trial arm assignment despite being collected by classroom teachers. Post-test ASBI scores were not blind to trial arm assignment. This was unavoidable given the need for a teacher familiar with the student to complete the questionnaire. We recognise this potential for bias in post-test ASBI scores as an unavoidable limitation to the findings related to this secondary outcome. The ASBI was chosen as it has been widely used as a measure of social-behavioural development and was relatively straightforward and practical to administer.

The final secondary outcome was the quality of the provision for language and social development as measured by a composite Environment Rating Scale (ERS) score based on items from the ECERS-3, ECERS-E, and SSTEW. Teachers in the treatment arm were trained to apply the ERS to their own classroom setting, which heightens awareness of the items that the ERS highlights in the treatment teachers' classrooms.

The items from each scale included in this composite measure are detailed in Appendix 1 of the Trial Protocol.[2] For transparency, total scores for the ECERS-3 and SSTEW, alongside the literacy subscale of the ECERS-E, are also reported in our evaluation. All scores used the standard scoring approach of summing question scores and dividing by the number of questions (minimum score = 1, maximum = 7). ERS scores for reception classes were collected by A+ Education staff prior to randomisation of schools to trial arms and were collected again in the autumn term 2017. The observers were blind to trial arm assignment.

The trial protocol stated that the Renfrew Bus Story Test would be piloted during pre-test as a possible additional post-test secondary analysis, with the final decision contingent on the cost of collection observed at pre-test. This measure was chosen to measure narrative language skills and mean-utterance-per-sentence. During pre-test it was administered to a small number of children in two schools. This was done after our primary language measures were administered, and children were given a break before the Bus Story was assessed. Ultimately, the cost of collecting the Bus Story Test was prohibitively high due to recording and transcription costs, and the decision was made to not alter the trial design to include this measure in our post-test language assessment battery.

[2] Note that one of the items listed in the trial protocol (Item 14 for the SSTEW: Assessing Language Development) was ultimately omitted from the composite measure. This is because in later work by the delivery team this item is no longer included—given its focus on assessment rather than on interactional strategies and the class environment. However, in practice the inclusion or exclusion of this item in the composite measure makes no substantive difference to the results.

Sample size

The sample size deemed necessary at the trial protocol stage followed standard EEF guidance on power (0.80), level of significance (0.05), and targeted minimum detectable effect size (MDES) (0.22). The specific assumptions the evaluation team made concerned test-retest correlation, the average number of children per school, the intracluster correlation coefficient (ICC) of the primary outcome measure, and child-level attrition.

For the language assessments used in this trial, we did not have information on test-retest correlation over the length of time this study would run for. However, based on test-retest correlation coefficients for language development assessments of similarly-aged children over similar lengths of time, we assumed a test-retest correlation coefficient of 0.50 (Sibieta et al., 2016). We assumed the average number of children per school to be 24, an estimate arrived at after discussion with the project delivery team. We assumed the ICC to be 0.20, which is a common, if conservative, estimate in education trials. Attrition was assumed to run at 15%, again a common education trial estimate. On this basis, maintaining a MDES of 0.22 (the target agreed by the EEF) required 117 schools to be enrolled in this trial (with no allowance for cluster-level attrition). This was rounded to 120 to adjust for the possibility of attrition at the school level.

Due to the high proportion of missing data for the pre-test, a decision was made when drafting the Statistical Analysis Plan to no longer include individual pupil pre-test scores but instead use school-level averages at pre-test. The test-retest correlation coefficient ultimately observed, given the school-level baseline average variable, was 0.28, but was estimated to be 0.50 in the sample size calculations detailed in the Evaluation Protocol. This lower than expected test-retest correlation impacts the MDES at the analysis stage.

Due to a mix of school- and student-level attrition (detailed in depth later in this report), the final sample size was 115 schools and 1,978 children. The ICC ultimately calculated was also considerably lower (0.13) than our original estimate (0.20). This meant the MDES at the point of analysis was 0.21 (see **Table 7**).

For the subgroup of children eligible for free school meals (FSM), at the time of writing the trial protocol we assumed four pupils per cluster would be eligible for FSM by the time they reached reception year (based on the 16.7% of students receiving FSM as determined by the average across all primary schools on the EduBase dataset). Given the prioritisation of schools with a larger proportion of FSM pupils, we had hoped these calculations would be conservative. The assumption of four FSM pupils per cluster gave an MDES of 0.29.

Ultimately the number of FSM-eligible children (as determined by the variable EVERFSM_6_P_SPR18 from the National Pupil Database) was 711, or seven per school, which meant at analysis the MDES for this subgroup was 0.23. Note that the FSM indicator was only requested from the National Pupil Database at the point of analysis, so at randomisation the MDES reported in **Table 7** still reflects our initial sample proportion assumption.

Randomisation

School recruitment was handled by the delivery team, which then handed on details of schools to BIT to deal with consent and data collection processes before randomisation was conducted by another member of the BIT research team. The project delivery team had no role in the randomisation process, other than to inform participating schools of their assignment.

Randomisation followed recruitment of schools, including the signing of Memoranda of Understanding (MoUs), distribution and receipt of opt-in consent forms by schools, provision to BIT of child-level data, and, in nearly all cases, the completion of baseline assessments.

Randomisation was stratified on the basis of school-level characteristics (proportion of FSM students and school location) to ensure balance between treatment and control groups (to be of equal or near-equal size). This was conducted using Stata and the randomisation code can be found in Appendix J. The randomisation followed a two-stage process:

1. The schools were stratified on the basis of FSM students (split across the median sample proportion as recorded in the Department for Education's Performance Tables) and location (split into West Midlands, Manchester, and Liverpool groups). This stratification, along with the two waves of randomisation described below, resulted in twelve blocks and schools were randomised within each block. The purpose of this blocking

was to improve cross-arm comparability of schools and also to increase precision of estimates. Block sizes are described in **Table 4** below.

2. A random number was generated within each block and the subsamples were split into two groups of equal size to ensure that school FSM proportion and location were balanced across trial arms.

Randomisation ultimately occurred in two waves due to baseline measure collection and recruitment time constraints. The two waves were randomised as described above. A total of 122 schools were recruited, with 62 randomised to treatment and 60 to control. Of the 122 schools, 96 were randomised in the first batch (on 25 November 2016) and 26 in the second batch (on 31 November 2016). Because the second batch contained far fewer schools and the location strata divided this already small batch into smaller blocks, the FSM median split did not always divide each location's schools evenly.

As timing did not permit schools, in many instances, to post consent forms to BIT before the randomisation date, BIT relied on schools to confirm consent forms had been distributed and provide a list of consenting children to satisfy the condition that consent had been obtained. Ultimately, two schools that provided lists of consenting children failed to deliver hard copies of consent forms to BIT. As such, two schools that were randomised were excluded from post-test data collection and analysis on the basis that no record of consent by children's parents was held.

Table 4 notes the final number of schools randomised by stratification variable.

Table 4: Randomisation strata by batch

Batch 1			
	West Midlands	Manchester	Liverpool
Below median FSM	29	11	7
Above median FSM	25	8	16 (15, after removing 1 school without consent forms)
Batch 2			
Below median FSM	7	3	5
Above median FSM	2	6 (5, after removing 1 school without consent forms)	3

Statistical analysis

As set out in the SAP, impacts were estimated for both the primary and secondary analysis using a least squares linear model with treatment arm indicators, stratum indicators (whether the school was above or below the median FSM proportion, school location, plus whether the school was randomised as part of the first or second batch) and a measure of prior attainment. Due to the issues with the collection of pre-test data discussed above, the primary analysis did not include the individual-level pre-test composite language skill score, instead controlling for the school average pre-test composite score.

As this is a school-level randomised controlled trial, inference was based on standard errors adjusted for school-level clustering using Stata's 'cluster' option to allow for correlation of pupil outcomes within schools.

The estimated impacts capture the effect of intention to treat (ITT) and are reported with 95% confidence intervals. Intra-cluster correlations (ICCs) are also reported. The model estimated for the primary analysis is:

$$\textbf{Equation 1: } Y_{ijt} = \alpha + \beta_1 Treat_j + \beta_2 Y_{jt-1} + \beta_3 \gamma_j + \varepsilon_{ijt}$$

where i are individuals and j are schools, Y_{ij} is our composite language skill score, $Y_{jt-1}$is the school average pre-test score, $Treat$ is our school-level treatment indicator, γ_j being a set of stratum dummies, where each stratum corresponds

to a unique combination of the stratification variables (including randomisation batch), and ε being an error term. Errors are clustered at school-level (j). Our primary intention to treat outcome is recovered from the estimate of β_1. This model was not altered according to the significance of any variables included (that is, all variables were retained in the model regardless of whether they are statistically significant) including the vector of blocking variables (γ_j).

We also explored each component of the composite score by conducting a factor analysis as a sensitivity analysis. The factor analysis posits that there is a latent factor describing language skills, with our four observed tests scores (counting the two APT scores separately) representing manifest measures of this underlying construct. Unlike the composite score proposed above, this allows the four measures to load on the common latent factor to differing degrees. We estimated the loadings of the factor on the four measures using an exploratory factor analysis principal factor approach, constraining there to be a single retained factor. From this model, we predicted values of the language skills latent factor using the regression scoring measure.

Impacts are estimated for secondary outcomes following the same model for the primary analysis specified above. Secondary analysis replaces Y_{ijt}, the composite language measure, with, separately, each of the secondary outcome measures specified earlier in this report (the four components of the composite language score: the ASBI, and the ERS ratings). Note that as ERS ratings are observed for one class per school, the models for the ERS outcomes are estimated at school level.

In line with the approach set out in the SAP, estimates are presented as effect sizes, calculated using the Hedges' g formula, expressing the estimated effect as represented by the regression coefficient relative to the total unadjusted outcome variance in the sample. Note, however, that the standardisation of the language score measure means that this is already in units of standard deviation for this outcome.

Given the number of secondary outcomes specified, we have applied the Benjamini-Hochberg procedure to adjust for multiple comparisons.

Impacts are also estimated using the same approach for the subgroup of pupils eligible for FSM (using the variables EVERFSM_6_P available from the NPD in line with EEF guidance). Impacts are also estimated for other subgroups specified in the protocol and SAP, namely for EAL pupils, those identified as having language difficulties based on their pre-test score, and by gender. To test whether there were differences for these subgroups, interaction terms were incorporated into the models. The analysis was run separately for the FSM subgroup as specified in the SAP.

The evaluation protocol specified that in the event that there was greater than 5% missing data at either cluster or individual level, or a significant difference in missingness between treatment and control arms, we would conduct further investigation into the mechanisms of missingness. The level of missingness for post-test data was 22% for both arms. Three schools withdrew following randomisation from control, and two from treatment, and a number of other pupils could not be tested in both arms due to absence, lack of assent, and language constraints. There is no difference in missingness between treatment and control arms; and at school (cluster) level, missingness is below the 5% threshold for further investigation. However, the level of missingness at individual level exceeded 5%. Given this difference, we pursue further exploratory sensitivity analysis as detailed in the SAP. We include an assessment of missing data at pupil level and investigate the extent to which baseline characteristics are correlated with non-response using linear regression and the same set of variables as detailed in our balance checks (the results are discussed within the additional analysis section later in this report). As a sensitivity analysis, we undertake multiple imputation of baseline data stratified on treatment condition as an additional robustness check. The model for imputation controls for participant school, gender, and free school meals status, and we use chaining to use baseline scores that were collected where we have partial cases.

The results of the additional analyses specified in the SAP are also presented. Sensitivity analyses included:

1. a variation of the primary model in which complete cases are included in the analysis, where analysis is restricted to pupils for whom both individual pre-test scores and post-test scores are available; here, the measure of prior attainment is the individual level pre-test rather than the school-level average;
2. a variation of the primary model for all pupils for which we have post-test data, and in which missing pre-test data is imputed using the approach described above; and
3. sensitivity analyses with the inclusion and exclusion of pupils for whom some, but not all four, language measures are available.

We also carry out an additional exploratory analysis, as specified in the SAP, of the extent to which change in ERS scores mediate the treatment effect. This aims to investigate whether any effect of the programme is working through changes in the learning environment (as proxied by composite ERS scores), as is hypothesised by the project's logic model. As noted in the SAP, we stress that this analysis is exploratory in nature since changes in learning environment are not randomly assigned. This is carried out using the following regression model:

Equation 2: $Y_{ijt} = \alpha + \beta Treat_j + \beta_2 Y_{ijt-1} + \beta'_3 \gamma_j + \beta_4 \Delta ERS_j + \beta_5(\Delta ERS_j * Treat_j) + \varepsilon_{ijt}$

where ΔERS is the change in composite ERS scores between those collected by the project team before randomisation ERS_{jt-1} and then collected contemporaneously with outcome measures ERS_{jt}, as appropriate. Our primary parameters of interest in this model are as follows: β_4 shows how outcomes vary with the general change in ERS in schools; β_5 shows the further variation associated with change in ERS in treatment schools. To explore whether it is change in particular aspects of the ERS that are associated with the change in outcome measure, we also conduct analyses that replace the composite ERS score with its component subscales.

Compliance

Compliance with the intervention is also discussed in the IPE section of this report. The following approach was agreed between the evaluation and delivery team, as documented in the SAP, in order to assess compliance for analysis of the primary outcome:

1) Did the pupil have a nursery teacher who attended over three training sessions?
2) Did the pupil have a reception teacher who attended over three training sessions?
3) Did the school engage with the intervention?

Engaging with the intervention meant that schools had:

a) used at least some elements of the URLEY tools and/or materials (for example, the ERS, TROLL, the Language Learning Principles, the interaction audit or other tools, or the action planning process);
b) attempted to implement changes within their classroom/s even though they may have faced challenges in doing this; and
c) made some attempt to introduce new staff to the approach, where staffing has changed.

Data was collected on engagement at both class and school level. Where there were differences between teachers or classes within a school, mentors were asked to make an overall judgement regarding schools as a whole, according to the following scale:

1—good, consistent engagement;
2—reasonable engagement (or, where this has been mixed, half or more of participating teachers have engaged);
3—some engagement (or, where this has been mixed, fewer than half of participating teachers have engaged); or
4—little or no engagement.

Scores of one or two were considered to be compliant with the intervention; scores of three or four were considered non-compliant. Mentors, who were part of the intervention team, assessed these engagement scores.

In practice, unfortunately it has not proved possible to establish whether pupils had a reception teacher who attended over three training sessions. While information is available on training attendance, unfortunately we do not have information linking pupils to the reception class they progressed to. For the analysis of compliance for the primary outcome, our measure of compliance is therefore based on attendance of the nursery teacher at three or more training sessions and the engagement score. These indicators are combined into a single indicator of compliance, with compliance defined as occurring when all criteria are met. Because we did not anticipate being unable to link pupils to reception classes, the SAP does not state this as an alternate definition of compliance; however, it was devised prior to running CACE analysis.

We use Complier Average Causal Effect (CACE) analysis to estimate intervention effects on treated children. We estimate the CACE using two stage least squares (2SLS) regression by estimating a (first stage) model of compliance using the binary measure of compliance described above. The predicted values from the first stage are then used in the estimation of a (reduced form) model of our outcome measure, Y_{ijt}. In other respects, the specification remains the same as the primary outcome ITT model. We conduct this analysis using the ivregress functionality of Stata to make necessary adjustments to standard errors (which will also be clustered at school level) due to the instrumental variables approach.

A compliance analysis is also undertaken, using the same approach described above, for the analysis of ERS scores. Here, the compliance criteria are defined as follows:

1) whether the teacher (of the class that is observed at post-test) had attended over three training sessions (that is, four or more);
2) the school's engagement score, (receiving a score of one or two considered to be complying with the intervention); and
3) whether this teacher was also the teacher observed at the pre-test.

Note that the analysis was not undertaken blind to randomisation. We do not feel that this biases the results because the evaluation team was not responsible for delivery of the intervention; thus, knowing the randomisation assignments would not impact or bias any interactions with the randomised schools.

Implementation and process evaluation

The overarching purpose of the process evaluation was to show how the URLEY intervention was delivered and implemented by settings, the factors that informed this, and any perceived impact it had upon children, staff, and classrooms. In addition, the process evaluation monitored the activity of the control group to establish what was done in the absence of the URLEY intervention. The process evaluation also aimed to bring greater clarity to the quantitative research findings and to understand the reasons behind them. It also gathered practitioners' views on how the intervention might be improved, to inform its future rollout.

The following research methods were used:

- visits to five treatment schools (referred to here as 'case study schools') conducted from November to December 2018; visits included interviews with nursery and reception teachers, teaching assistants, EYFS leads, a tour of the EYFS setting, and in some cases an interview with the head or the assistant headteacher;[3]
- telephone interviews with the four mentors;
- detailed end-of-project survey of treatment schools piloted and administered from February to May 2018—sent to all participating nursery and reception teachers in each treatment setting;
- survey of control schools administered from May to July 2018—distributed to all schools in the control group at the time that research assistants were visiting schools to conduct the post-test language assessments;[4]
- observations of three training days (third and fourth sessions), the third session was attended in two locations (Warrington and Birmingham); and
- a review of programme materials.

All evaluation activities were carried out by the Behavioural Insights Team (BIT) and the National Institute of Economic and Social Research (NIESR), with support from the delivery team. The delivery team also provided the raw data showing the mentors' engagement ratings for each school. This was based on mentors' assessment on a number of

[3] Due to staff illness, interviews were done over the phone for one of the case study schools.

[4] The control survey was administered slightly later than the end-of-project treatment survey in order to maximise response rates by coinciding with the language assessments. The surveys are, however, treated as comparable in the analysis, but the findings should be interpreted in light of this.

indicators[5] ranked from one to three as well as a summary rating. The delivery team also provided data on teacher attendance at training and ERS scores.

The end-of-project survey of treatment schools was completed by 56 respondents out of 144 (39% response rate). At a school level, the survey was completed by 39 out of the 60 treatment schools (65% response rate). The end-of-project survey was very detailed and took about 20 minutes to complete and included a number of open-ended questions. As such, the survey gathered detailed qualitative data. We undertook a number of robustness checks to explore to what extent the sample could be considered representative of the full set of intervention schools. These show that respondents are very similar to non-respondents in terms of their engagement and which mentor they had. Clearly though, non-respondents could be atypical in other ways not captured by these indicators, compared to the broader sample of teachers involved in the study. In an attempt to explore this, we therefore followed-up with non-respondents at a later stage and asked them to answer the final question of the survey, which asked for their overall recommendation of the programme. In total, this question was answered by 82 respondents out of 144 (57% response rate). At a school level, the final question was answered by 50 out of the 60 treatment schools (83% response rate). The answers of those who did not respond to the main survey were similar to the responses of those who did.

A brief survey was sent out to control schools at the same time as the research assistants conducted the language assessments for the post-test. The questions were designed to understand how the nursery environment of control school settings differed from those of the intervention group. In total, 37 control schools returned the survey, which represents over 60% of the total control group. Steps were also taken to ensure that the five case study schools included a variety of delivery contexts. That is, treatment schools were selected that differed by Ofsted rating, proportion of FSM pupils, and geographical location. We also sampled two of the ten schools that had no respondents to the end-of-project evaluation survey. However, the case study schools are a relatively small proportion of the treatment group as a whole. The findings may not, therefore, necessarily reflect the views of the wider population of treatment and control schools. Nevertheless, taken together with the survey data, we believe the qualitative data collected through the visits provides useful insights into the range and diversity of views, and the experience of participants in the URLEY intervention. The findings of the process evaluation should be considered with these strengths and limitations in mind.

We visited case study schools relatively late in implementation—in November and December 2018—around two years after URLEY started in September 2016. This meant that the evaluation team had obtained the findings of schools' progress (as assessed by their ERS scores); this enabled us to sample schools that had achieved either low, medium, or high progress, and the relatively late visits also allowed the evaluation team to explore to what extent URLEY had become embedded in the treatment schools and in teacher practices.

Costs

Data on costs was collected from A+ Education, the University of Oxford, and the schools themselves.

Delivery partner data

The delivery partner provided information on the cost of training and time costs relating to the one-to-one mentor provision and training times. The cost of training was the actual cost that would be incurred by schools were it not subsidised by the EEF. The time costs we have provided are for six days' training, but it is important to note that A+ Education have subsequently adapted the model and now only offer five days of training. However, we have reported six days for the purpose of this evaluation as that is the intervention which the trial schools experienced.

School cost survey

A cost survey (in paper form) was sent out to schools as part of the outcomes data collection at post-test. Just over half (52%) of the treatment schools responded to the survey, 31 schools in total. Schools were asked to report on the costs of the intervention to the school—the materials they had purchased, travel and subsistence, the cost of covering staff

[5] There were a total of nine indicators: vision for URLEY; mindset and motivation; knowledge, understanding and skills; reflection and analysis; tuning in; development of practice; leadership and team engagement; the QI process; and growing independence.

at training, the cost of sharing and cascading the intervention to other staff, and the cost of any new physical materials purchased, based on feedback from the mentor and the ERS, to improve the classroom environment.

We also asked schools to report on time spent embedding the intervention in their school—time at training, time spent being mentored, as well as time to plan to incorporate the recommendations at their setting. Staff were also asked to report on the time taken to cascade the intervention to their teaching assistants, and, in some cases, to the whole school.

We used data from this survey, alongside data from the delivery partner, A+ Education, to calculate the financial and time costs outlined later in this report.

Timeline

The table below describes the timeline for the research activities and overall study.

Table 5: Timeline

Date	Activity
April 2016–November 2016	School recruitment led by delivery team
25 November 2016	Batch 1 randomisation by BIT
31 November 2016	Batch 2 randomisation by BIT
5 October–12 December 2016	Pre-tests conducted by BIT RAs
22 March-30 June 2017	Observation of training days by evaluation team
September 2017–May 2018	Roll-out of the intervention
1 May 2018–31 July 2018	Post-tests conducted by BIT RAs
February–May 2018	Survey of treatment schools by evaluation team
May–July 2018	Survey of control schools by evaluation team
November–December 2018	Case study visits and mentor interviews by BIT/NIESR
August 2018–30 August 2019	Analysis and reporting conducted by evaluation team

Impact evaluation

Participant flow including losses and exclusions

As noted, 122 schools were ultimately recruited: the original target overshot as in the weeks immediately prior to the randomisation deadline a large number of schools had to be approached in order to ensure the target of 120 was met.

At randomisation, the MDES achieved differed from that predicted in the trial protocol (see **Table 7** for MDES levels throughout the trial). Firstly, at 21, the average number of children per school was lower than the 24 anticipated. This was due to several quite small schools being recruited (eleven schools had fewer than ten children enrolled). Secondly, we relied on schools reporting that opt-in consent forms had been distributed and returned in order to count them as having fulfilled the selection criteria and being able to proceed to randomisation. However, two schools—despite stating that this occurred and therefore being randomised—failed, ultimately, to send BIT their opt-in consent forms. These schools should not have been randomised into the trial as, in the absence of consent forms, pupil data could not be held or processed.[6] As such, the MDES at randomisation was 0.22, reflecting 120 schools and a total of 2,535 children.

An unforeseen issue in this trial was the number of pupils who would be unable to be assessed, both at pre- and post-test. There were several reasons why some children did not complete assessments:

- Several schools in the trial had a high number of children from migrant backgrounds with low levels of proficiency in English. This meant these pupils were unable to engage with the test, leaving a smaller number of children in these schools to be assessed.
- Some children were absent across multiple days, and therefore not present on the days when assessments were collected.
- Some children did not consent to participate in the assessments at the time of testing.
- Some children had learning and/or physical impairments that meant they were unable to participate in the assessments.
- Some children had left their schools and could not be located and tested at their new schools (post-test only).

The trial protocol, while setting out selection criteria for schools, did not set out selection criteria for pupils that would have seen pupils excluded from the study on the basis of not having a pre-test collected (for any of the reasons above). As a result, some children are included in the study without having a pre-test assessment (see **Table 6**). Pupils with missing post-test data, for any reason, are counted as part of attrition from the study.

[6] In Figure 1 these two schools are excluded from the number of schools randomised, as it is considered that these schools should not have been randomised as part of the trial.

Table 6: Extent of missingness in data collection

Data	n/N (Proportion of sample missing)	Reason(s)
Composite pre-test scores	Pupil level: 459/2535 (18%) C = 222; T = 237 School level: 0/120	Pupils were absent during testing, could not be tested on the day due to time constraints, or faced language constraints that limited their ability to be tested. In some cases, assent was not gained as the child did not want to take the test.
Composite post-test scores	Pupil level: 557/2535 (22%) C = 276; T = 281 School level: 5/120 C = 3; T = 2	For five schools (pupil n = 129), it was not possible to collect post-test data as they withdrew from the trial post randomisation. The remaining pupils (n = 428) could not be tested due to absence from school, having moved schools, or could not be tested for other reasons on the day.
Unmatched NPD data	31/2535 (1%)	The NPD could not find a match for pupil UPNS.
Total ASBI pre-test scores	1843/2535 (27.3%)	22/120 schools did not return pre-test ASBI score sheets for children.
Total ASBI post-test scores	1632/2535 (35.6%)	23/120 schools did not return post-test ASBI score sheets for children.

In light of almost one fifth of the sample lacking a pre-test score, it was agreed in discussions with the EEF that a school-level mean of pre-test results be used in the primary analysis rather than individual child pre-test results as specified in the trial protocol. Given this and the school and pupil attrition observed, the MDES at analysis should have fallen, but the intracluster correlation observed was significantly lower than predicted, meaning the final MDES is actually lower than that estimated at trial protocol stage (see **Table 7**).

Figure 1: Participant flow diagram

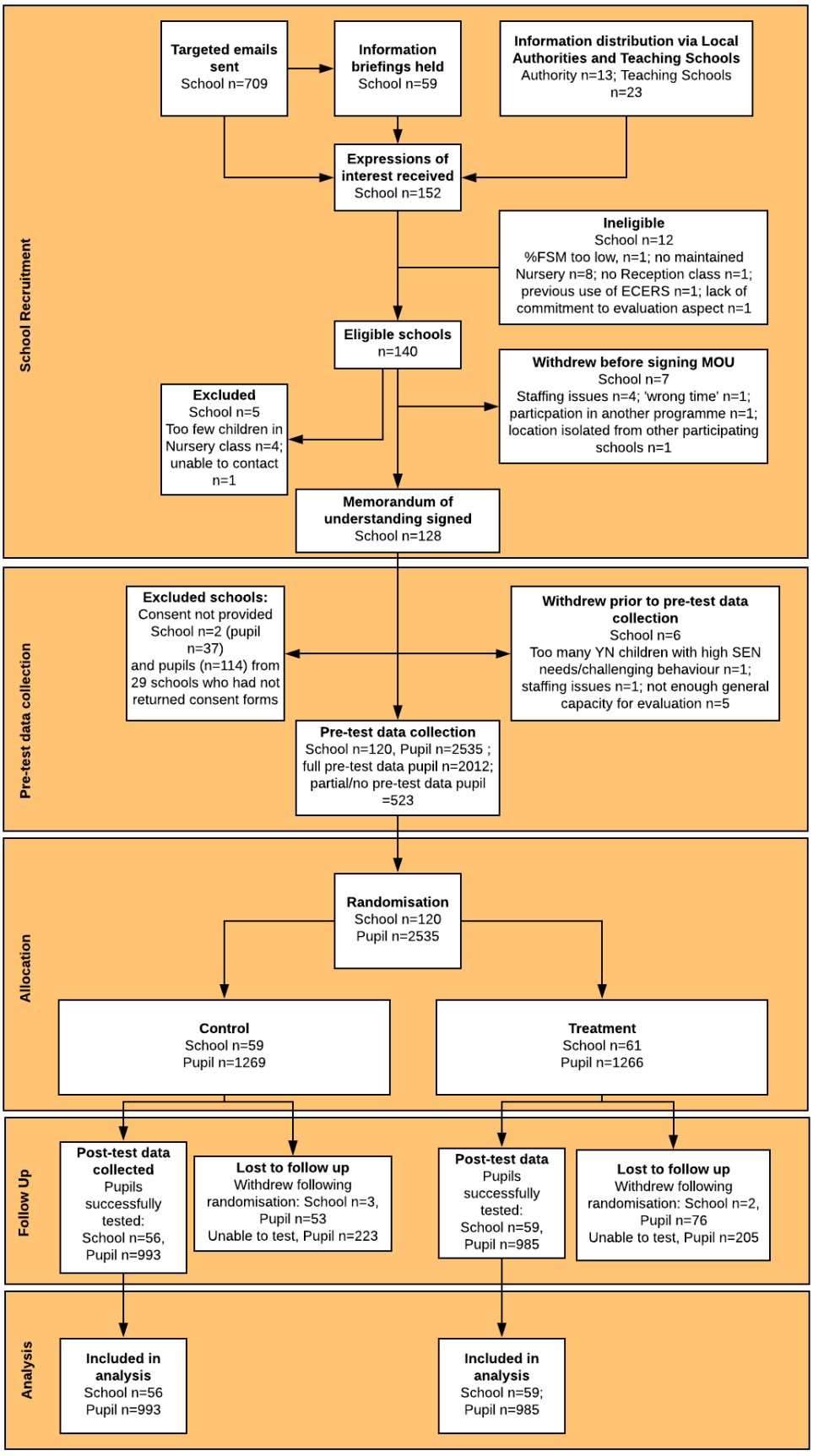


Table 7: Minimum detectable effect size at different stages

		Protocol		Randomisation		Analysis	
		Overall	FSM	Overall	FSM	Overall	FSM
MDES		0.22	0.29	0.22	0.30	0.21	0.23
Pre-test/post-test correlations	level 1 (pupil)	0.5	0.5	0.5	0.5	0.28	0.57
	level 2 (class)	-	-	-	-	-	-
	level 3 (school)	-	-	-	-	-	-
Intracluster correlations (ICCs)	level 2 (class)	-	-	-	-	-	-
	level 3 (school)	0.20	0.20	0.20	0.20	0.13	0.13
Alpha		0.05	0.05	0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8	0.8	0.8
One-sided or two-sided?		2	2	2	2	2	2
Average cluster size		24	4	21	4	17	6
Number of schools	intervention	60	60	61	61	59	54
	control	60	60	59	59	56	54
	total	120	120	120	120	115	108
Number of pupils	intervention	1440	240	1,288	215	985	322
	control	1440	240	1,246	208	993	389
	total	2880	480	2,535	488	1,978	711
Attrition predicted/observed		15%	15%	15%	15%	22%	23%

Attrition

The sample size calculations produced for the trial protocol assumed 15% attrition at the pupil level. Ultimately attrition in this trial was slightly higher, at 22% of the sample at randomisation. There were two contributing factors to this higher than anticipated attrition rate.

- Firstly, five schools withdrew from the trial in the period between randomisation and post-testing (see Figure 1 for further detail). These schools withdrew because they did not want to commit to post-intervention data collection.
- Secondly, and most significantly, it was not possible to administer our assessments to all children for whom parental consent was granted for reasons described above.

Concerns around sample size were first identified at the pre-test point as some pupils could not complete the assessments. The rate at which these issues prevented assessments from being collected was higher than predicted, with the underestimation a product of both few studies of this kind having been conducted previously (and thus little guidance on this issue existing) and the recruitment of schools in areas with higher than average proportions of children with English as an Additional Language (EAL).

Pupil and school characteristics

Table 8 presents school and pupil-level characteristics for the intervention and control group at the point of randomisation. On most of the characteristics we are able to observe, the sample was balanced at baseline. The distribution of schools by Ofsted rating was similar for both trial arms, with the majority rated either ‘good’ or ‘outstanding’. The distribution by school type was also similar in the two arms. The majority of participating schools

were primary schools (93% of schools in both the treatment and control groups), with most of the remainder being infant schools, and one all-through school in the control group.[7] Almost all schools, regardless of treatment arm, were located in urban areas. There were also no substantive differences by trial arm in school size or in the composition of pupils within schools (based on the percentage of pupils in the school eligible for FSM, the percentage for whom English is an additional language or the percentage with special educational needs). However, when we consider characteristics at pupil level, while the sample is balanced in respect of gender and age, a smaller percentage of pupils allocated to the treatment group were eligible for FSM compared with the control group, while there were a higher percentage of EAL pupils in the treatment group compared with the control group.

Due to issues with the completion of the pre-test, as discussed above, the primary analysis uses a school-level average for the pre-test measure; this was similar in both trial arms, but slightly higher among the intervention group (effect size = 0.11). We also report the average individual pre-test scores in **Table 8** below and present histograms showing the distribution of these scores in Appendix E. In addition to the composite language measure, scores for the four component measures are presented. Mean scores were generally similar in the treatment and control groups. The greatest difference is observed for the BPVS subscale, with a higher mean score among pupils in the treatment group compared with the control group (effect size of 0.09).

Table 8: Baseline comparison

School level (categorical)	Intervention group		Control group	
	n/N (missing)	Count (%)	n/N (missing)	Count (%)
Ofsted overall effectiveness:[1]				
Outstanding	10/60 (1)	10 (16.7%)	6/56 (3)	6 (10.7%)
Good	42/60 (1)	42 (70.0%)	45/56 (3)	45 (80.4%)
Requires improvement	7/60 (1)	7 (11.7%)	2/56 (3)	2 (3.6%)
Inadequate	1/60 (1)	1 (1.7%)	3/56 (3)	3 (5.4%)
School type:[2]				
Academy converter	8/61 (0)	8 (13.1%)	8/59 (0)	8 (13.6%)
Academy sponsor led	3/61 (0)	3 (4.9%)	7/59 (0)	7 (11.9%)
Community school	31/61 (0)	31 (50.8%)	22/59 (0)	22 (37.3%)
Foundation school	0/61 (0)	0 (0%)	1/59 (0)	1 (1.7%)
Voluntary aided school	16/61 (0)	16 (26.2%)	18/59 (0)	18 (30.5%)
Voluntary controlled school	3/61 (0)	3 (4.9%)	3/59 (0)	3 (5.1%)
In urban area[3]	59/61 (0)	59 (96.7%)	59/59 (0)	59 (100%)

[7] This is identified on the basis of the age of pupils within the schools, as provided in the DfE Performance Tables.

School level (continuous)[2]	n (missing)	Mean (SD)	n (missing)	Mean (SD)
Number of pupils on roll	61 (0)	366.6 (154.7)	59 (0)	350.5 (155.4)
% pupils eligible for FSM	61 (0)	23.7 (10.4)	59 (0)	25.7 (13.3)
% pupils ever eligible for FSM	61 (0)	40.3 (14.0)	59 (0)	42.5 (17.4)
% pupils with EAL	61 (0)	26.5 (28.3)	59 (0)	23.4 (25.7)
% pupils with SEN support	61 (0)	15.9 (7.6)	59 (0)	14.4 (6.7)
Pre-test composite language score: school level average	61 (0)	0.04 (0.51)	59 (0)	-0.01 (0.42)
Pupil level (categorical)[4]	n/N (missing)	Count (%)	n/N (missing)	Count (%)
Eligible for FSM	322/1253 (16)	322 (25.7%)	389/1251 (31)	389 (31.1%)
EAL/unclassified	348/1253 (0)	348 (27.8%)	262/1251 (0)	262 (20.9%)
Male	639/1253 (0)	639 (51.0%)	629/1251 (0)	629 (50.3%)

Pupil level (continuous)	n (missing)	Mean (SD)	n (missing)	Mean (SD)	Effect Size
Pre-test composite language score	1003 (263)	0.04 (1.02)	1009 (260)	-0.01 (0.99)	0.05
Pre-test: BPVS	1029 (237)	0.06 (1.02)	1046 (223)	-0.03 (0.99)	0.09
Pre-test: RAPT Information	1015 (251)	0.00 (1.01)	1028 (241)	0.01 (0.99)	-0.01
Pre-test: RAPT grammar	1015 (251)	0.03 (1.01)	1028 (241)	-0.02 (1.01)	0.05
Pre-test: CELF	1011 (255)	0.05 (1.00)	1017 (252)	-0.01 (1.00)	0.06
Age at September 2016 (months)[4]	1253 (0)	43.534 (3.516)	1251 (0)	43.472 (3.616)	

Notes and sources:

1. Ofsted inspection ratings as at 31 December 2016.
2. As reported in in DfE Performance Tables, 2017.
3. As reported in Edubase, schools located in ‘urban city and town’ or ‘urban major conurbation’, data accessed April 2019. All other participating schools were located in ‘rural town and fringe’ areas
4. An additional 15 cases in the intervention group and 16 cases in the control group could not be matched to the NPD.

All other characteristics are as reported in the NPD extracts provided for this project.

Table 9 presents the same set of school and pupil characteristics for the analysis sample. This provides us with an insight into whether attrition may have led to, or accentuated, any imbalance in the sample, although this does not appear to be the case. The sample remains balanced in respect of the same characteristics that were observed to be balanced at baseline. As at the point of randomisation, the treatment group has a higher percentage of EAL pupils and

a lower percentage of FSM pupils, compared with the control group. However, these differences in terms of FSM and EAL were already present at randomisation and do not appear to have been influenced by attrition. We undertook an additional analysis including the percentage of FSM pupils and EAL pupils as additional controls to test the sensitivity of our results.

The average composite language pre-test scores for the analysis sample were also fairly similar in the two trial arms, whether this is assessed at individual level (effect size = 0.07) or in terms of the school-level average (effect size = 0.10), with slightly higher mean scores in the intervention group. There was no difference in the mean RAPT information score by trial arm, while for the BPVS and CELF, mean scores were slightly higher in the treatment group (effect size = 0.09) and to some extent also for the RAPT grammar score (effect size = 0.04). Histograms presenting the distribution of pre-test and post-test scores are shown in Appendix E.

Table 9: Comparison—analysis sample

School level (categorical)	Intervention group		Control group	
	n/N (missing)	Count (%)	n/N (missing)	Count (%)
Ofsted overall effectiveness:[1]				
Outstanding	9/58 (1)	9 (15.5%)	6/53 (3)	6 (11.3%)
Good	42/58 (1)	42 (72.4%)	42/53 (3)	42 (79.2%)
Requires improvement	6/58 (1)	6 (10.3%)	2/53 (3)	2 (3.8%)
Inadequate	1/58 (1)	1 (1.7%)	3/53 (3)	3 (5.7%)
School type:[2]				
Academy converter	8/59 (0)	8 (13.6%)	7/56 (0)	7 (12.5%)
Academy sponsor led	3/59 (0)	3 (5.1%)	7/56 (0)	7 (12.5%)
Community school	31/59 (0)	31 (52.5%)	20/56 (0)	20 (35.7%)
Foundation school	0/59 (0)	0 (0%)	1/56 (0)	1 (1.8%)
Voluntary aided school	14/59 (0)	14 (23.7%)	18/56 (0)	18 (32.1%)
Voluntary controlled school	3/59 (0)	3 (5.1%)	3/56 (0)	3 (5.4%)
In urban area[3]	57/59 (0)	57 (96.6%)	56/56 (0)	56 (100%)
School level (continuous)[2]	n (missing)	Mean (SD)	n (missing)	Mean (SD)
Number of pupils on roll	59 (0)	363.8 (156.4)	56 (0)	357.7 (156.1)
% pupils eligible for FSM	59 (0)	23.8 (10.5)	56 (0)	25.4 (13.4)
% pupils ever eligible for FSM	59 (0)	40.3 (14.3)	56 (0)	42.2 (17.4)

% pupils with EAL	59 (0)	26.4 (28.7)	56 (0)	24.2 (26.1)	
% pupils with SEN support	59 (0)	15.9 (7.7)	56 (0)	14.4 (6.9)	
Pre-test composite language score: school-level average	59 (0)	0.05 (0.52)	56 (0)	-0.01 (0.43)	
Pupil level (categorical)[4]	n/N (missing)	Count (%)	n/N (missing)	Count (%)	
Eligible for FSM	247/985 (9)	247 (25.1%)	299/991 (21)	299 (30.2%)	
EAL/unclassified	275/985 (0)	275 (27.9%)	202/991(0)	202 (20.4%)	
Male	505/985 (0)	505 (51.3%)	485/991 (0)	485 (48.9%)	
Pupil level (continuous)	n (missing)	Mean (SD)	n (missing)	Mean (SD)	Effect Size
Pre-test composite language score	803 (182)	0.08 (1.01)	802 (191)	0.02 (1.00)	0.066
Pre-test: BPVS	825 (160)	0.09 (1.03)	832 (161)	0.00 (0.98)	0.09
Pre-test: RAPT Information	814 (171)	0.04 (1.00)	819 (174)	0.04 (0.99)	0.00
Pre-test: RAPT grammar	814 (171)	0.06 (1.00)	819 (174)	0.02 (1.00)	0.04
Pre-test: CELF	809 (176)	0.09 (1.00)	806 (187)	-0.00 (1.02)	0.09
Age at September 2016 (months)[4]	985 (0)	43.582 (3.473)	991 (0)	43.477 (3.583)	

Notes and sources:

1. Ofsted inspection ratings as at 31 December 2016.
2. As reported in DfE Performance Tables, 2017.
3. As reported in Edubase, schools located in ‘urban city and town’ or ‘urban major conurbation’, data accessed April 2019. All other participating schools were located in ‘rural town and fringe’ areas.
4. An additional two cases in the control group could not be matched to the NPD.

All other characteristics are as reported in the NPD extracts provided for this project.

Table 10 presents absolute standardised differences for those characteristics specified in the statistical analysis plan, for the analysis sample. These are consistent with the findings described above, indicating some degree of imbalance by trial arm in terms of the percentage of pupils eligible for FSM and the percentage of pupils with EAL.

Table 10: Absolute standardised differences for selected characteristics—analysis sample

	Intervention group mean	Control group mean	Absolute standardised difference
% male pupils	51.3	48.9	4.66
% pupils ever eligible for FSM	25.1	30.2	12.3
% schools rated outstanding or good by Ofsted	87.9	90.6	8.4
% pupils with EAL	27.9	20.4	17.7
Age in months	43.6	43.5	2.96

Note: all characteristics presented in the table are pupil-level characteristics, with the exception of Ofsted ratings, which are school-level.

Outcomes and analysis

Primary analysis

Table 11 summarises the results of the primary analysis. Comparison of the mean scores for the treatment and control groups indicates a slightly lower mean on the composite language measure (the primary outcome) in the intervention group (-0.016) compared with the mean for the control group (0.016). The distribution of scores across treatment and control groups are presented in Figure 2.
The analysis indicates a slightly negative, but non-statistically-significant impact of the programme on the primary outcome measure (the magnitude of the effect is equivalent to one month's less progress). Given the outcome measure is standardised, the impact estimates are already in units of standard deviation. For completeness however, and in line with the SAP, the parameters for the effect size calculation are reported in **Table 12**.

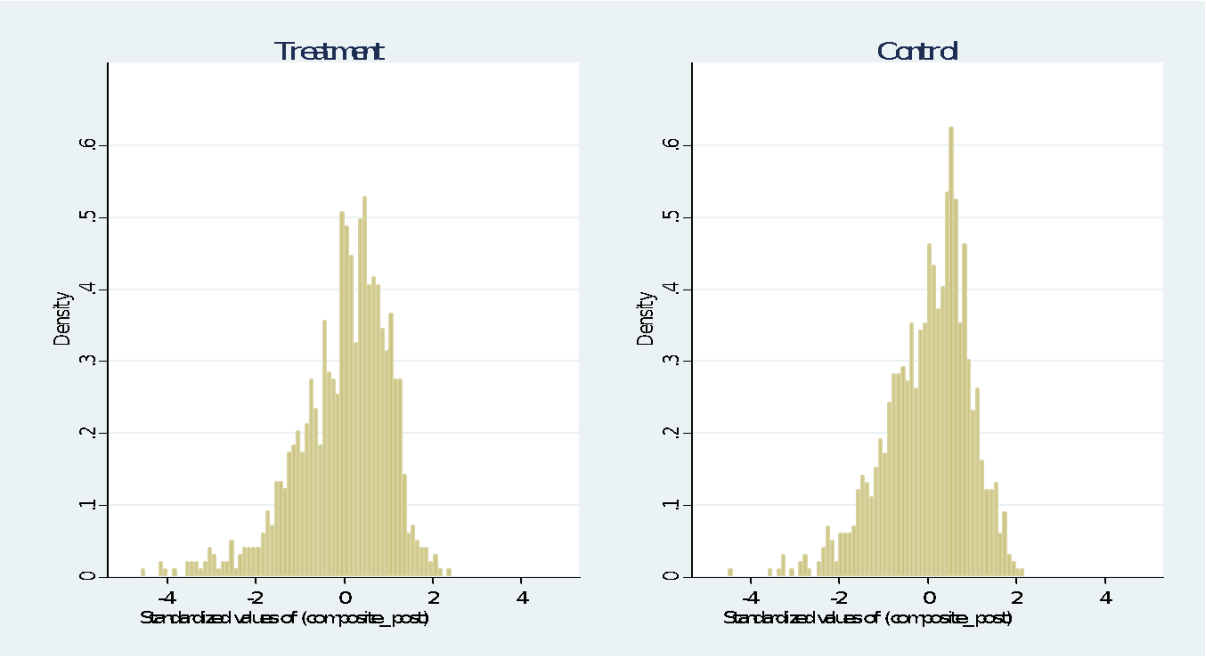
Table 11: Primary analysis

	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)	p-value
Composite language measure	985	-0.016 (-0.081, 0.049)	993	0.016 (-0.043, 0.075)	1,978 (985; 993)	-0.08 (-0.19, 0.03)	0.15

Table 12: Effect size estimation

Outcome	Unadjusted differences in means	Adjusted differences in means	Intervention group		Control group		Pooled variance	Population variance (if available)
			n (missing)	Variance of outcome	n (missing)	Variance of outcome		
Composite language measure	-0.032	-0.081	985	1.088	993	0.913	1.000	-

Figure 2: Histograms of composite language score by trial arm



As an exploratory sensitivity analysis on the primary outcome measure, we conducted a factor analysis by estimating the loadings of the factor on the four language measures that comprise the composite language score. The factors loadings for each component are very similar. Using these, we then predicted the values of the language skills latent factor using the regression scoring measure. These results, along with the primary regression model on the factor that was calculated, can be found in Appendix J. This makes no substantive difference to the results of the primary analysis.

Secondary Analysis

As specified in the trial protocol, we consider the four individual measures that form the composite language score as secondary outcomes.[8] The results for these four scales are presented in **Table 13**. This shows small, negative, but not statistically significant impacts on the BPVS score, the RAPT information score, and RAPT grammar score; the effect sizes correspond to one months' less learning. There is some indication of a statistically significant, negative impact of 0.122 (corresponding to two months' less learning) on the CELF Sentence Structure test (-0.231, -0.014). As specified in the SAP, we account for the 35 comparisons in the secondary/additional/exploratory analysis by applying the Benjamini-Hochberg procedure, which requires adjusting p-value thresholds when judging statistical significance. Assuming a false discovery rate of 0.05, the smallest observed p-value would need to be smaller than 0.05/35 = 0.001 in order to be considered statistically significant. In the case of the CELF score, we no longer consider the observed effect to be statistically significant.

[8] For each model, the analysis controls for the school-level pre-test average for the respective outcome measure.

Table 13: Secondary outcomes—language outcome subscales

	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)	p-value
BPVS	985	0.009 (-0.071, 0.052)	993	-0.009 (-0.071, 0.052)	1978 (985; 993)	-0.058 (-0.149, 0.033)	0.214
RAPT Info	985	-0.021 (-0.038, 0.081)	993	0.021 (-0.038, 0.081)	1978 (985; 993)	-0.044 (-0.168, 0.081)	0.490
RAPT Grammar	985	-0.018 (-0.042, 0.078)	993	0.018 (-0.042, 0.078)	1978 (985; 993)	-0.049 (-0.177, 0.079)	0.454
CELF	985	-0.023 (-0.038, 0.083)	993	0.023 (-0.038, 0.083)	1978 (985; 993)	-0.122 (-0.231, -0.014)	0.029

As specified in the SAP, impacts on children's scores on the ASBI are also considered as secondary outcomes. **Table 14** reports the results, which show no statistically significant impact on either the total ASBI score[9], or on any of the four component scales. In terms of magnitude, the results indicate very small negative or almost zero effect (the effect sizes for the overall ASBI score, comply, disrupt[10] and prosocial subscales correspond to zero months' additional progress, and the express subscale effect size corresponds to 1 months' less learning). Histograms for each of the ASBI follow-up scores are presented in Appendix F. These show highly skewed distributions which lead us to have some concerns about floor/ceiling effects. This suggests that it would therefore be difficult for any intervention, however effective, to influence these scores for this sample.

Table 14: Secondary outcomes—ASBI

	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)	p-value
Total ASBI	703	84.494 (83.975, 85.013)	688	84.855 (84.332, 85.377)	1391 (703; 688)	-0.040 (-0.184, 0.104)	0.588
Express	723	36.352 (36.080, 36.624)	703	36.594 (36.306, 36.882)	1426 (723; 703)	-0.065 (-0.214, 0.085)	0.399
Comply	720	28.654 (28.444, 28.865)	750	28.648 (28.444, 28.852)	1470 (720; 750)	0.005 (-0.136, 0.147)	0.940
Disrupt	731	19.318 (19.153, 19.483)	753	19.404 (19.243, 19.566)	1484 (731; 753)	0.003 (-0.177, 0.183)	0.973
Prosocial	708	65.122 (64.691, 65.554)	694	65.330 (64.888, 65.772)	1402 (708; 694)	-0.024 (-0.170, 0.123)	0.752

The final secondary outcome considered is the quality of provision. **Table 15** presents present this in relation to language and social development, as measured by ERS scores. Baseline and follow-up observations were collected

[9] The total ASBI score was calculated by combining the scores of the express, comply, and disrupt subscales. The prosocial scale is not included in this calculation because it consists of components that already contribute to the express, comply and disrupt subscales.
[10] We have coded the disrupt scale to reflect the same pattern as the express and comply scales, where a higher score represents a better outcome.

for 60 intervention schools and 57 control schools (based on observation of one reception class in each participating school). As specified in the SAP, four ERS measures are reported: a composite measure based on a selection of items from the ECERS-3, ECERS-E, and SSTEW (with the items to be used specified in the trial protocol), as well as the total scores on each of the ECERS-3 and SSTEW, and the literacy subscale of ECERS-E.

For all four measures, higher mean scores were observed in the intervention group compared with the control group when observed at follow-up. Histograms for each of the measures are presented in Appendix G. Overall, the results of this analysis indicate a positive and statistically significant impact of the programme on each of the four reported measures of quality of provision.

Table 15: Secondary outcomes—ERS scores

	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)	p-value
Composite ERS	60 (1)	3.45 (3.21, 3.69)	57 (2)	2.89 (2.62, 3.15)	117 (60; 57)	0.69 (0.41, 0.96)	0.00
ECERS-3	60 (1)	3.10 (2.93, 3.27)	57 (2)	2.80 (2.60, 3.00)	117 (60; 57)	0.58 (0.31, 0.86)	0.00
ECERS-E (literacy subscale)	60 (1)	3.66 (3.47, 3.84)	57 (2)	3.24 (3.00, 3.47)	117 (60; 57)	0.62 (0.32, 0.91)	0.00
SSTEW	60 (1)	3.06 (2.81, 3.31)	57 (2)	2.64 (2.38, 2.90)	117 (60; 57)	0.52 (0.26, 0.78)	0.00

Subgroup analysis for primary and secondary outcomes—composite, four language scores, and ASBI

FSM subgroup

Table 16 reports the results of estimating the primary model for the FSM subgroup, with **Table 17** presenting the underlying parameters for the effect size calculation. The estimated effect size for this subgroup is of the same magnitude as for the full sample, and again is not statistically significant.

Table 16: Impact on primary outcome—pupils eligible for FSM

	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)	p-value
Composite language measure	247 (0)	-0.081[11] (-0.201, 0.040)	299 (0)	-0.053 (-0.155, 0.050)	546 (247; 299)	-0.08 (-0.27, 0.11)	0.395

[11] A z-score expresses the number of standard deviations from the mean a data point is. A negative z-score expresses standard deviations below the mean.

Table 17: Effect size estimation—pupils eligible for FSM

Outcome	Unadjusted differences in means	Adjusted differences in means	Intervention group		Control group		Pooled variance	Population variance (if available)
			n (missing)	Variance of outcome	n (missing)	Variance of outcome		
Composite language measure	-0.028	-0.077	247	0.934	299	0.814	0.868	-

Table 18 reports the underlying parameters for the effect size calculation of the language outcome subscales amongst pupils eligible for FSM. Similarly, this shows no statistically significant impact on scores on the BPVS, or on either the RAPT information score or RAPT grammar score. There is some indication of a statistically significant, negative impact of 0.224 on the CELF Sentence Structure test (-0.416, -0.033). This impact is slightly more pronounced than in the overall analysis that includes all pupils. However, when adjusting for multiple comparisons by applying the Benjamini-Hochberg procedure as described above, we find that the results for the CELF are no longer reliable.

Table 18: Secondary outcomes—language outcome subscales, pupils eligible for FSM

	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)	p-value
BPVS	247	-0.048 (-0.171, 0.074)	299	-0.122 (-0.228, -0.016)	546 (247; 299)	0.023 (-0.165, 0.211)	0.809
RAPT Info	247	-0.051 (-0.174, 0.072)	299	-0.021 (-0.126, 0.084)	546 (247; 299)	-0.052 (-0.227, 0.123)	0.562
RAPT Grammar	247	-0.060 (-0.183, 0.062)	299	-0.041 (-0.151, 0.069)	546 (247; 299)	-0.032 (-0.226, 0.162)	0.748
CELF	247	-0.104 (-0.225, 0.018)	299	0.013 (-0.092, 0.118)	546 (247; 299)	-0.224 (-0.416, -0.033)	0.024

In addition, we consider the ASBI score and subscores amongst pupils eligible for FSM. **Table 19** reports the results, which show no statistically significant impact on either the total ASBI score, or on any of the four component scales.

Table 19: Secondary outcomes—ASBI, pupils eligible for FSM

	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)	p-value
Total ASBI	155	83.766 (82.646, 84.885)	178	83.345 (82.356, 84.333)	333 (155; 178)	0.020 (-0.212, 0.252)	0.904
Express	165	36.347 (35.808, 36.887)	184	36.164 (35.609, 36.720)	349 (165; 184)	0.015 (-0.216, 0.247)	0.896
Comply	164	28.302 (27.813, 28.791)	194	28.354 (27.961, 28.747)	358 (164; 194)	-0.068 (-0.300, 0.163)	0.565
Disrupt	167	19.116 (18.774, 19.458)	197	18.826 (18.485, 19.168)	364 (167; 197)	0.072 (-0.163, 0.307)	0.531
Prosocial	159	64.650 (63.718, 65.581)	179	64.518 (63.690, 65.347)	338 (159; 179)	0.026 (-0.211, 0.262)	0.833

EAL pupils

As specified in the trial protocol, we also explore whether there are differential effects for EAL pupils. Pupils in the EAL subgroup, on average, performed significantly worse on the composite language measure and component scales than pupils not in the EAL subgroup (**Table 20**). To test whether the treatment had a particularly strong effect on pupils in these subgroups we interact EAL status with treatment allocation. The results show no evidence that the intervention had a differential effect for EAL pupils.

Table 20: Composite language measure and component scales—EAL subgroup

	Composite language score	BPVS	RAPT Info	RAPT Grammar	CELF
Treatment	-0.010 (0.062)	0.044 (0.052)	-0.028 (0.069)	-0.000 (0.074)	-0.056 (0.059)
EAL	-0.554*** (0.104)	-0.468*** (0.100)	-0.483*** (0.099)	-0.464*** (0.099)	-0.406*** (0.106)
Treatment x EAL	-0.041 (0.133)	-0.151 (0.128)	0.079 (0.138)	-0.011 (0.144)	-0.065 (0.121)
N	1,976	1,976	1,976	1,976	1,976

Note: Each column shows selected coefficients from a regression of the outcome on Treatment, EAL, Treatment*EAL, pre-test score, and blocking dummies of strata used in randomisation. School-level clustered standard errors are reported in parentheses. Statistical significance indicated as follows: * p < 0.05; ** p < 0.01; *** p < 0.001.

In contrast, **Table 21** reports that there is no significant difference between the EAL subgroup and non-EAL subgroup in terms of the total ASBI score and subscales. The intervention did not have a significant effect on ASBI outcomes when comparing across pupils in the EAL subgroup.

Table 21: Total ASBI and component scales—EAL subgroup

	Total ASBI	Express	Comply	Disrupt	Prosocial
Treatment	-0.268 (0.623)	-0.004 (0.303)	0.003 (0.242)	-0.123 (0.269)	0.093 (0.503)
EAL	0.306 (0.820)	-0.119 (0.414)	0.116 (0.347)	0.309 (0.229)	0.044 (0.710)
Treatment x EAL	-0.070 (1.038)	-0.712 (0.589)	0.039 (0.450)	-0.416 (0.307)	-0.698 (0.921)
N	1390	1425	1469	1483	1401

Note: Each column shows selected coefficients from a regression of the outcome on Treatment, EAL, Treatment*EAL, pre-test score, and blocking dummies of strata used in randomisation. School-level clustered standard errors are reported in parentheses. Statistical significance indicated as follows: * p < 0.05; ** p < 0.01; *** p < 0.001.

Pupils with language difficulties

The analysis sample contained 691 children with language difficulties (324 in treatment and 367 in control).[12] These pupils, on average, performed significantly worse on the composite language measure, component scales and total ASBI, express, comply, and prosocial scores than pupils not classified as having language difficulties (**Tables 22** and **23**). Pupils with language difficulties did not perform significantly worse on the disrupt score. The inclusion of interaction terms did not provide any evidence of the treatment having differential effects on pupils with language difficulties compared with those that did not; however, this characteristic was not included in the balance checks.

[12] Defined as those who scored in the bottom 15% of BPVS age-standardised scores during the pre-test (that is, one standard deviation below the mean of the normed population).

Table 22: Composite language measure and component scales—language difficulties subgroup

	Composite language score	BPVS	RAPT Info	RAPT Grammar	CELF
Treatment	-0.0700 (0.060)	-0.0226 (0.052)	-0.0851 (0.070)	-0.0897 (0.073)	-0.0324 (0.064)
Language Difficulties	-0.863*** (0.067)	-0.856*** (0.071)	-0.626*** (0.065)	-0.719*** (0.068)	-0.611*** (0.078)
Treatment x LD	0.0119 (0.104)	-0.0153 (0.100)	0.0996 (0.103)	0.102 (0.112)	-0.165 (0.115)
N	1656	1656	1656	1656	1656

Note: Each column shows selected coefficients from a regression of the outcome on Treatment, Language Difficulties, Treatment*Language Difficulties, pre-test score and blocking dummies of strata used in randomisation. School-level clustered standard errors are reported in parentheses. Statistical significance indicated as follows: * p < 0.05; ** p < 0.01; *** p < 0.001.

Table 23 : Total ASBI and component scales—language difficulties subgroup

	Total ASBI	Express	Comply	Disrupt	Prosocial
Treatment	-0.523 (0.556)	-0.336 (0.258)	-0.0768 (0.205)	-0.062 (0.254)	-0.318 (0.411)
Language Difficulties	-2.269*** (0.539)	-1.599*** (0.328)	-0.687** (0.205)	-0.055 (0.165)	-2.262*** (0.453)
Treatment x LD	0.006 (0.920)	-0.234 (0.460	-0.111 (0.387)	0.084 (0.288)	-0.287 (0.731)
N	1157	1187	1208	1222	1165

Note: Each column shows selected coefficients from a regression of the outcome on Treatment, Language Difficulties, Treatment*Language Difficulties, pre-test score and blocking dummies of strata used in randomisation. School-level clustered standard errors are reported in parentheses. Statistical significance indicated as follows: * p < 0.05; ** p < 0.01; *** p < 0.001.

Gender

As demonstrated in **Tables 24** and **25**, female pupils performed, on average, better than male pupils on the composite language score, the RAPT information score, and the RAPT grammar score. For the CELF score, we do find some evidence of differential effects of the treatment by gender; however, adjusting for multiple comparisons renders this effect unreliable as the p-value of this interaction term (0.006) is greater than 0.001, the threshold for reliable effects within 35 comparisons and a false discovery rate of 5%. On average, females scored higher on the total ASBI score and all of the subscales, compared to males. We find no evidence, however, of differential effects of the treatment on ASBI scores by gender.

Table 24: Composite language measure and component scales—female subgroup

	Composite language score	BPVS	RAPT Info	RAPT Grammar	CELF
Treatment	-0.154* (0.068)	-0.114 (0.064)	-0.0907 (0.078)	-0.0853 (0.080)	-0.228*** (0.062)
Female	0.117* (0.059)	0.00481 (0.058)	0.153* (0.067)	0.151* (0.061)	0.0736 (0.053)
Treatment x Female	0.154 (0.085)	0.110 (0.082)	0.105 (0.092)	0.0838 (0.088)	0.214** (0.078)
N	1976	1976	1976	1976	1976

Note: Each column shows selected coefficients from a regression of the outcome on Treatment, Female, Treatment*Female, pre-test score and blocking dummies of strata used in randomisation. School-level clustered standard errors are reported in parentheses. Statistical significance indicated as follows: * p < 0.05; ** p < 0.01; *** p < 0.001.

Table 25: Total ASBI and component scales—female subgroup

	Total ASBI	Express	Comply	Disrupt	Prosocial
Treatment	0.051 (0.688)	-0.154 (0.399)	0.0787 (0.309)	0.212 (0.235)	0.0632 (0.617)
Female	2.737** (0.471)	1.003** (0.347)	1.202*** (0.207)	0.719*** (0.157)	2.165*** (0.449)
Treatment x Female	-0.605 (0.720)	-0.169 (0.461)	-0.0691 (0.291)	-0.377 (0.210)	-0.359 (0.645)
N	1390	1425	1469	1483	1401

Note: Each column shows selected coefficients from a regression of the outcome on Treatment, Female, Treatment*Female, pre-test score and blocking dummies of strata used in randomisation. School-level clustered standard errors are reported in parentheses. Statistical significance indicated as follows: * p < 0.05; ** p < 0.01; *** p < 0.001.

Additional Analysis

In this section we explore the additional analyses specified in the SAP: we run the primary analysis including only those with pre- and post-scores (replacing the school-level average pre-test with the individual pre-test scores), imputing pre-scores where these are missing, and, thirdly, for those with pre- and post-scores, analysing each subscore separately. We find that all additional specifications of our analysis are consistent with the main analysis. The results of this additional analysis are reported in Appendix H.

In Appendix H, we also present the results of running an additional analysis that also controls for whether pupils are eligible for FSM or have EAL status. This analysis was not pre-specified in the SAP, but we ran this in response to the observed imbalance between treatment and control groups on these two characteristics. In this model, the magnitude of the effect size is reduced to -0.02, which corresponds to no additional (or less) progress. We discuss this further in the interpretation section.

As specified in the SAP, and as described earlier by Equation 2 in the Statistical Analysis section of this report, we also undertook an additional exploratory analysis to consider the extent to which any change in ERS scores might mediate the treatment effect. The original aim of such an analysis was to explore whether any effect of the programme appeared to be working through changes in the learning environment (as proxied by the ERS scores). This analysis is conducted solely for our primary outcome measure, the composite language score. Arguably, given our primary analysis indicates no statistically significant impact; there is perhaps less merit in undertaking this additional exploratory analysis;

nevertheless, we report the results for completeness in Appendix I. This analysis was conducted using both the composite ERS score and also repeated for each of the three subscales: ECERS-3, ECERS-E (literacy subscale), and the SSTEW. In all models, neither the coefficient for the estimated change in outcome associated with the change in the ERS score in treatment schools, nor the coefficient on the change in outcome unexplained by the change in the ERS score in treatment schools, is statistically significant.

The SAP specified that in the event of more than 5% missing data further investigation into missing data would be undertaken. Around one fifth (22%) of post-test data was missing for both trial arms. The results of our assessment of this missing data at pupil level are reported in further detail in Appendix H. We began by regressing an indicator of whether the post-test data is missing on the same variables used in our primary analysis specification. We then extended this to additionally include the pupil-level characteristics used in our balance checks. The results show no significant difference in missingness between treatment and control arms, and no significant differences according to the other observed characteristics. We also checked for the presence of interactions between treatment allocation and other characteristics. However, no statistically significant interactions were found.

Compliance

Finally, in this section we present the results of analyses assessing compliance. As described earlier in the report, we constructed a binary indicator of compliance for analysis of the primary outcome measure. On this basis, just under half of pupils were in schools considered to be 'compliant' with the intervention (N = 479 pupils). The estimated effect size for the CACE estimate is equal to -0.16 (p-value = 0.16). Although this effect appears larger in magnitude than the ITT estimate, equivalent to two months' less progress, it is still not statistically significant.

A similar analysis was conducted for the ERS outcome measures. Again, we construct a binary measure of compliance, based on the criteria specified in the SAP. It is worth noting that on this basis only 20 of the intervention settings were deemed to be fully compliant with the intervention. The effect of the intervention does appear to be larger among 'compliers'; the CACE estimate of the effect (coefficient of 1.90, p-value 0.00) is around three times the magnitude of the ITT estimate.

For both analyses, it is worth noting that this approach implicitly attributes all the effect to compliers. That is, the estimated intention to treat effect presented earlier is interpreted as arising purely from pupils in treatment schools that satisfied the compliance definition. Given the nature of the compliance indicator, some caution should be exercised in interpretation; arguably, it is perhaps more plausible that treatment schools that fell short of the compliance definition would just experience a smaller effect, rather than no effect. Such schools, after all, may have implemented the intervention to some extent, just not the extent required to fully meet the criteria for compliance. Indeed, the findings presented in the implementation and process evaluation section later in this report are consistent with some schools having partially implemented and engaged with the programme.

Cost

Schools participating in the intervention received six days of training from January 2017 to November 2017 for both nursery and reception teachers.[13] Each school individually received, on average, 17.5 hours of mentoring, a combination of face to face, and online support.

The more substantial financial costs that the schools incurred to deliver the intervention related to training, the cost of staff cover, and, in some cases, the cost of adding new elements to the early years setting. The majority of these costs were accrued in the first year, with very minimal additional costs in subsequent years. See **Table 26**.

We assumed 25 pupils, instead of the number of treated pupils, as the trial was opt-in; we lost a number of pupils so the ITT does not reflect actual class size. Therefore we take an average of 25, as this follows the estimate used by the EEF in their toolkit (Education Endowment Foundation, 2018). It is worth noting that in such CPD programmes, multiple year-groups of children would potentially benefit from teachers that received training, and so actual per pupil costs may

[13] In some cases, schools sent more than two teachers if they had a multi-form entry nursery and reception. In one case, a teacher was replaced by a higher level teaching assistant.

be lower. The cost here is based on a single cohort that was evaluated, rather than all children that would have been taught by teachers receiving the programme during the project. See **Table 27.**

Table 26: Individual school cost of delivering URLEY

Item	Detail	Type of cost	Average cost	Total cost 3 years	Total cost per pupil per year over 3 years
Materials	Laminating, handouts, printing, photocopying new displays	Running costs	£19.35	£58.05	£0.77
Expenses	Staff travel, subsistence	Start-up costs	£16.34	£16.34	£0.22
Training fees	Training	Training costs	£ 3,400.00	£3,400.00	£45.33
Physical environment	Changes to classroom	Start-up costs	£410.88	£410.88	£5.48
Total			£3,846.57	£3,885.27	£51.80

(Source: A+ Education/Oxford University; Teacher Cost Survey)

Table 27: Cumulative costs of URLEY (assuming delivery over three years)

	Year 1	Year 2	Year 3
URLEY	£3,846.57	£3,865.92	£3,885.27

As outlined above, the intervention becomes better value for money when delivered to a higher number of pupils (assuming that staff do not leave and need retraining). The costs are 'low' by EEF definitions—under £5,000 per school, or £200 per pupil.

Table 28: Start-up vs training costs in Year 1

Per school	£	Per Pupil[14]
Training cost	3,400	136.00
Start-up costs	446.36	17.85
Running costs	19.35	0.49
Total cost Y1	3,858.52	154.34

(Source: A+ Education/Oxford University; School Cost Survey)

[14] Cost survey based on 25 pupils.

Outliers

If schools had to make substantive changes to their early years setting, for example, buying a new learning area or particular equipment, they incurred substantial costs—as much as £2,000. This applied to roughly 10% of our sample that returned the cost survey.

Staff time

The most substantial cost to schools is staff time as the intervention requires six whole days of training for both nursery and reception teachers (40 hours per teacher), and 17.5 hours mentor contact time with the school. This time cost is magnified as the majority of cost survey respondents highlighted that they paid for cover for the training days, which at least two teachers attended, only one school reporting they covered the days internally. On average, schools needed 14.3 days of supply cover. The range of reported hours to embed and cascade the intervention, and cover staff absence, was therefore vast, ranging from 1 hour to 40 days. Schools also reported taking time to cascade the learning of the intervention and planning for the changes in weekly or semi-regular short meetings (range one to six hours).

Implementation and process evaluation

The findings presented in this section are based on the qualitative data collected from case studies, surveys, observations, and mentor interviews. The team made visits to five case study schools that were selected to include a variety of delivery contexts and progress. The end-of-project survey of treatment schools was completed by 56 respondents out of 144 (39% response rate). At the school level, the survey was completed by 39 out of the 60 treatment schools (65% response rate). The control school survey, which was filled out by one early years staff member from each school, was completed by 34 schools (60% of non-participating schools).

While steps were taken to ensure that the treatment schools visited as part of fieldwork included a variety of delivery contexts, the number that participated in this part of the research was very small compared to the larger treatment group. As such, the findings may not necessarily reflect the views of the wider group of treatment and control schools. Nevertheless, taken together with the survey data, we believe the qualitative data collected through the visits provide useful insights into the range and diversity of views, and the experience of participants in the URLEY intervention. The findings of the process evaluation should be considered with these strengths and limitations in mind.

Implementation

Preparedness for delivering the intervention

Overall, survey respondents and case study interviewees said the training and support had equipped them well to deliver the intervention. **Table 29** below shows that 91% of survey respondents said the URLEY training sessions, materials, and mentoring supported them ‘quite a lot’ or ‘very much’ to implement the intervention in their class and school.

Table 29: To what extent did the URLEY training sessions, materials and mentoring support you to implement URLEY in your class and school? (N = 56)

	Number (%)
Very much	23 (41%)
Quite a lot	28 (50%)
A little	5 (9%)
Not very much	0 (0%)
Not at all	0 (0%)

Survey respondents were also asked to give a rating for each element of the programme to reflect how important these were in supporting their knowledge base and/or improvements to practice. The responses to these showed that teachers attached great importance to all elements of the URLEY programme that were prompted for. Though all elements were rated highly, some were rated higher than others. **Table 30** shows the elements in descending order, according to the sum of people who responded ‘quite a lot’ and ‘very much’.

Table 30: What do you think of the different elements of the URLEY programme? Please give a rating below to each element to reflect how important they were in supporting your knowledge base and/or improvements to practice. (N = 56). Numbers are reported as sum of people who responded positively, that is, 'quite a lot' or 'very much'.

	Number (%)
Support from your URLEY mentor	53 (95%)
Time for reflection on your children/practice during training sessions	53 (95%)
Your physical pack/folder of resources	53 (95%)
The URLEY Language Learning Principles	52 (93%)
Time to share and discuss with colleagues during training sessions	52 (93%)
The training days (overall)	52 (93%)
The TROLL and other tools/guidance for capturing your children's language progress	50 (89%)
The Environment Rating Scales (ECERS-3, SSTEW and ECERS-E)	49 (88%)
The other tools and strategies in your folder (e.g. OWL, Incredible Teachers, interaction audit, conversation strategies)	49(88%)
Average	48 (86%)
Watching DVD clips of practice during the training sessions	45 (80%)
Online copies of URLEY documents	44 (79%)
Online DVD clips	38 (68%)
The research readings and time provided to read these	35 (63%)

The survey and case study visits also asked open-ended questions about what was rated most highly and least highly, giving a perspective on the views behind the headline numbers (**Table 30**). In the subsections below, the case study findings are also integrated to give a detailed account of how participants viewed the different elements of the URLEY programme. It should be said that some respondents made the point that all elements were of high quality and that they worked well together:

> *'I genuinely cannot think of an element which I would rate 'least', it's a jigsaw and each component complements the other.'* Nursery Teacher, Survey respondent 16.

Mentoring

The URLEY mentor played a critical role in ensuring effective implementation of the intervention. The survey findings indicated that the mentoring was seen as one of the most positive parts of the URLEY intervention: 95% of respondents answered that it supported their knowledge-base and/or improvements to their practice by 'quite a lot' or 'very much' (see **Table 30** above). Similarly, the vast majority of teachers found the mentoring process useful (**Table 31**) and stated that they had a positive relationship with their mentor (**Table 32**).

Table 31: Overall, how useful did you find the mentoring? (N = 56)

	Face-to-face mentoring	Distance mentoring (phone, Skype, email)
	Number and percentage	Number and percentage
Very helpful	39 (70%)	11 (20%)
Quite helpful	14 (25%	27 (48%)
Neutral	3 (5%)	15 (27%)
Not very helpful	0 (0%)	1 (2%)
Not at all helpful	0 (0%)	2 (4%)

Table 32: Please rate your relationship with your mentor. (N = 56)

	Number and percentage
Very positive	45 (80%)
Quite positive	11 (20%)
Neutral	0 (0%)
Quite negative	0 (0%)
Very negative	0 (0%)

Across the survey and case studies, teachers highlighted the value mentors added. The role was perceived to be so effective because the support was tailored to each school, enabling the mentor to work with the individual strengths and weaknesses of each school.

> '*[E]very school was unique and every teacher was unique ... while there was some common elements and some things that we knew that were going to be important cover ... the mentoring that was needed and the approach that was needed did vary.*' Mentor 1, Interview.

> *'She was very responsive to the needs of our school and would cater discussion around the challenges which we faced. Her advice was timely and specific rather than generic.'* Nursery Teacher, Survey respondent 16.

Early years staff, senior leaders, and mentors described that they gave support, or were supported, in two core ways:

- structural support was offered to tackle barriers that prevented the intervention being fully adopted, either by cascading the training to other staff or by working with the SLT to help create more internal school support for the programme; and
- pedagogical support was provided—mentors worked in class, modelling particular concepts, supporting with planning or making changes to the classroom environment, and, in some cases, working individually with children.

Although the principal purpose of the mentor was 'to support the schools and the staff in implementing the ideas and the principles that they were learning about in the training … [and] implementing the training within … everyday practice' (Mentor 2), often there were structural barriers which the mentor tackled, such as getting senior leaders engaged, cascading intervention to support staff, or training new staff. Participants cited—during interviews, the survey, and in the observations—that having this external voice was valuable to overcome barriers to engagement, both from teaching assistants and senior leaders:

> *'Some schools had quite a good system for cascading ... where they always had URLEY on the agenda. For others, I supported with that on my visits. There was one school in particular where ... they organised my visit for a team meeting so myself and the head would work together on the content of it and we'd work together on delivering that.*' Interview, Mentor 3.

> *'Teacher explains to others on her table that her mentor has been really useful to help her make the case for prioritising URLEY to management.'* Researcher fieldnotes, CPD Session 4.

> *'[W]e were struggling to get the TA's on board ... She worked so hard to get them on board. By the end of the year she had a meeting with all of them, she did a really good session.'* Teacher 10, EYFS Lead, School 5.

In the survey, mentors were invariably described as knowledgeable and highly experienced practitioners. Some teachers said it had been useful to have an 'external outside eye' and 'an expert contact' visiting their school. Respondents said that mentors had acquired detailed knowledge about the individual setting and the cohort of children. This enabled them to provide advice that was specific to the school and cohort—and feasible to implement in the sense that it accounted for the barriers and strains of classroom life.

Within the classroom, the mentors provided a range of pedagogical support. The most commonly cited were planning, modelling good practice, environmental changes, and child-level support.

In the survey, teachers said their mentor had been very helpful and supportive, particularly around planning and embedding URLEY. Mentors had provided ideas and guidance on the URLEY programme and its techniques and had recommended strategies on how to move things forward and helped to put things into action. This was achieved by making suggestions of changes, working together to develop plans in the setting, giving feedback on planning documents, and helping to refine next steps. Two teachers in the case study schools mentioned that the mentor used the ERS to assess them and provide feedback:

> *'The ECERS is what she rated us on. It was more that we use the [Language Learning Principles] and she gave us feedback on it, using the ECERS.'* Teacher 1, School 1.

Mentoring sessions had also allowed time for reflection and discussion on the implementation of the URLEY programme and on teaching and school practices more broadly. The sessions gave teachers an opportunity to discuss what had worked and what issues teachers faced, and mentors were able to give feedback and engage in a constructive dialogue on improvements to teaching practice and on the classroom environment.

Modelling good practice was also cited widely by teachers and teaching assistants as particularly valuable. Survey respondents highlighted that it had been beneficial to observe their mentor demonstrating the URLEY techniques. This enabled them to see URLEY in practice, and some said it had been particularly useful to see URLEY techniques applied to their own classroom with their own children. Some specifically mentioned that it was useful to have the mentor walk through materials to capture the children's language progress, such as TROLL, and that it had been useful to observe mentors reinforcing new words by modelling them in a range of contexts.

> *'She demonstrated the commentary method with a child who rarely speaks, but it was amazing to see him start to open up to her, a complete stranger, with the right approach.'* Reception Teacher, Survey respondent 4.

This was echoed by teachers in the case studies who found it valuable for both themselves and their support staff.

> *'She was really useful … she'd model lessons and we'd watch her, or she'd watch us, the way we develop language. She'd observe us and give us feedback.'* Teacher 4 Case Study School 2.

As well as classroom practice, modelling observation using the ERS also enabled teachers to 'tune into their practice': 'She did observations with me, paired observations. She was excellent' (Teacher 9, EYFS Lead, School 5). Teachers cited that this helped them conduct more in-depth observations and had the secondary benefit of overcoming some of the negative connotations associated with observations and Ofsted:

> *'[Before] I was more casually observing and not really going in depth with the observations but now I found that I'm really unpicking them and getting a lot out of them.'* Teacher 3, Case Study School 2.

This practice was valued so much by some teachers that some mentors decided to roll out the practice across all the schools they were mentoring. Several examples were also cited where the ERS observation templates had been distributed beyond the EYFS setting:

> '*I did do a joint observation with the head and the head of early years and we walked round and picked out various things that they could see because the Head wanted to know how to use them because she was going to incorporate them into learning walks.*' Interview, Mentor 4.

For some teachers, the mentors played a large role in instigating changes to the early years environment, both through evolution of set ups and innovation with new areas, such as the Home Corner. As well as changing the environment, interviewees felt the mentor also worked with the teachers and support staff to reinforce the language of the different learning environments:

> *'She just asked us to go into an area, the Market, and just work with a group of children and record everything that was being said and that was happening, so we did. Afterwards she gave us a quick recap of what we'd*

> *done and as we were doing it, she came round to listen and she just told us exactly what we'd done and how we could have made it better. She explained how we could have improved the language and the bits that we did that was really, really good.'* Teaching Assistant 4, School 3.

Mentors also supported teachers with particular children, to focus on the 'reluctant' or 'invisible' children. This was particularly helpful for teachers struggling with non-verbal children, to allow the child intensive one-to-one support. It also helped the teacher and their staff to better understand the techniques of working with these children:

> *'[For] the children who had no speech at all, she would commentate on what they were doing and that was really useful for me and my TA at the time to see that, because we weren't doing that completely.'* Teacher 1, School 1.

Mentors also cited numerous examples of how they had worked with individual children and found it a particularly rewarding component of their role—and the results were so tangible:

> '*There was a little boy who had been in school at that point for about three months and he hadn't yet spoken to the staff. Again, we used some modelling, and I used the technique of commentary with him and commentary with play with me imitating some of his play and commenting on it. Within about five minutes he was talking to me, and she was like, "Oh my goodness, I've been trying to do that for three months." So that kind of had a difference.*
> *The next time I went into that school, that little boy as, soon as he saw me, came over to me again to talk to me. His teacher said it's made such a difference to him, but I could see the difference in him.'* Mentor 2.

However, it is worth noting that working with individual children on a one-to-one basis, although valuable, may not be practical for all teachers. For instance, one teacher preferred other types of support:

> *'If she would have asked us prior to the day what we would like to work on, to come in with new fresh ideas rather than sitting in the areas. She came once and she sat in the area with one child for the whole time and we were, like, "It's not really giving us an insight to anything else." She just sat in the play dough area she was in, and she was just talking to this child. I thought, well, "I could have done that. I could have sat there for that long talking to that child, but I haven't got an hour to be with one child, I've got a full class of 30 children to look after. I can't spend one hour with one child."'* Teacher 6, School 3.

This may be an example of a miscommunication between the mentor and the teacher, who had misaligned ideas of the purpose of the mentoring for that day. Given the time constraints, individual child work may seem wasteful to some, but improved communication—particularly rationalising the reasons behind this type of work—may help teachers as well as the pupils themselves.

As **Table 31** shows, the distance mentoring was generally considered less helpful than the face-to-face mentoring though it was still viewed in a predominantly positive light. The distance mentoring consisted of emails, phone calls, and Skype calls. A few respondents said they preferred emails as it was more flexible and did not require scheduling a timeslot. One respondent said she had not been aware that distance mentoring was available and had not made use of it. Generally, when respondents commented on distance mentoring, they said it had still been helpful and valuable, and some said that mentors had been very approachable and always available, either online or by phone, to answer questions when needed. This was echoed in the case studies where teachers valued the flexibility and availability of mentors:

> *'It was good as well not just for her to come in but also to have a telephone contact as well and that was just as useful and flexible enough without having to plan too much time out of the timetable to say, "Right, we'll sit down together then."'* Teacher 8, School 4.

Some less-positive aspects of the mentoring process were also mentioned in open-ended answers to the survey. Some teachers said that time constraints had been an issue as it had been difficult to arrange times that suited all school staff and mentors, and that it had been difficult to find cover and fit the visits around other school priorities. A few respondents said the scheduling process ideally could be improved but were unsure how this could be achieved in practice. Others highlighted that the changes that they had adapted to their environment and teaching had not been a result of advice from their mentor but from their own ideas. One interviewed school felt that other mentors had provided more support to other schools than it had received as its mentor has not been present at the training sessions:

> *'I personally don't feel that we had an awful lot from her to be honest. I think everything that we tried to do we took from our own initiative from the training days and as a team tried to do it. When we did go to the training days, other schools were talking about their experience; other mentors seemed to be very engaged with their teachers and knew what their school settings were like. You know, people on the training were staying behind afterwards to talk to their mentors, but we didn't really get that opportunity because ours was never really there.'* Teacher 10, School 5.

These respondents also argued that the guidance had been too broad and that they had needed more practical ideas and support. One respondent said that mentoring sessions could have been more mentor-led rather than asking teachers what they wanted to focus on. This sentiment was echoed by mentors, who, in hindsight, would have opted for a more structured support system:

> *'What we learnt was that schools needed ... [things to be] a little bit more structured in terms of saying actually, after setting one, we think these are the key things that you really need to focus on.'* Mentor 1.

This more structured approach would have been valuable to help focus the teachers who are working under multiple conflicting priorities, and help make a case to their SLT about the specific requirements resulting from the intervention such as additional resources or focus areas:

> *'I think having that kind of a structure to the mentoring, which initially we felt that, as professional people, they would be able to work out what [was] needed. But we didn't understand that not all of them were able to do that and work under huge pressure is because of this.'* Mentor 2.

However, many of the positive responses seemed to suggest that it was exactly this tailored and reflective approach that was valued by a majority of respondents. There is a balance to strike between a structured and open-ended approach to mentoring, and this may need to be decided on a case-by-case basis, depending on the current strengths and weaknesses of the school.

Training

There were five full-day training sessions held between January and June 2016, which all teaching staff within nursery and reception classes at the setting were expected to attend. Due to the high turnover of staff between 2015/2016 and 2016/2017, an additional induction day was held in the autumn term of 2016 to support staff new to the intervention.

The training comprised several components that the interviewees and survey respondents found valuable: pedagogical practice, theory, and networking. These elements were complemented by periods of reflection and planning as teams. The respondents who commented positively about the training said it was helpful, informative, and stimulating; specifically, some practitioners said the training had struck an appropriate balance between instruction and time for sharing and reflecting on practice. The time to speak to other schools and practitioners about implementation was considered particularly valuable, especially the opportunity to discuss what they were doing and how it was working.

The survey responses indicated that the training had been important in supporting practitioners' knowledge base and/or improvements to their practice, with 93% answering 'quite a lot' or 'very much.' In particular, the videos were mentioned as a tool to both learn about how to question children, but also how to observe other staff.

> *'The videos that they used within the day, that was good to see that the way you observe ... how to properly get in-depth observations.'* Teacher 4, School 2.

This meant that the teachers felt they had improved their own practice, but also knew how to better support and develop other staff through critical observation. As well as modelling observation, questioning was also highlighted as particularly valuable.

> *'One of the things was not just asking a question, [but] trying to involve yourself in the child's play, begin a conversation, and then weave your question in; not immediately "What are you painting?" ... giving them time to answer and not putting them under pressure … stepping back and just giving them a little bit of time or you carry on a conversation and then going back to them ... [that] was invaluable really.'* Teacher 10, School 5.

The practical knowledge conveyed to teachers was facilitated by the expertise of the trainers who often provided real examples of practice they had done themselves or had seen others do. These were observed by researchers to be both

positive examples, but also ones where a practitioner overcame a challenge. This was felt to complement the video stimuli and help ground the practice within everyday situations.

As well as practical pedagogy, learning the theory underpinning the URLEY programme was felt to be an essential component, as this senior leader summarised:

> *'From my experience of being a head and teaching in the past … people can put good practice in—what appears to be good practice—but unless you have a real understanding about why you're doing something, it doesn't work. The emphasis seems to get lost or it gets filtered. Unless your teachers have that expertise, you don't have that consistency across the department ...You can only really improve something if you really deeply understand why you're doing something.'* Headteacher, School 3.

Some survey respondents said they had benefited from gaining a more theoretical understanding of language development as well as the importance of environment and adult-child interaction. In particular, the 3 Million Word Gap video was referenced several times by teachers in both the survey and case studies to be insightful and motivating as well as underlining the importance of the programme.

The opportunity to share ideas was valued as teachers could pick up ideas for their own classroom or setting. This was compounded by a sense that the opportunity to build a network specifically around literacy was rare, and less developed than Special Educational Needs Coordinators networks, for example. The benefit of sharing ideas and examples of language was highlighted in the case study interviews and survey responses:

> *'This other school brought in new vocab cards and straight away you could see using it in our areas, with vocab that we want to focus on, and that was valuable.'* Teacher 6, School 3.

The time allotted for participants to think and reflect upon their own practice was highlighted—in both the case studies and the surveys—as useful. Although some cited that the training was too long at points, the time to reflect upon their own practice felt to be useful, particularly as teachers often stated that they had difficulty setting aside time to reflect during normal day-to-day activities.

> *'It was more of deeper thinking that we were encouraged to do—[something] you don't get a chance to think about once you start your practice and you're involved in everyday … work. You don't really … have much time and that's why it was nice to come out of the classroom and have a think about your personal interaction and your own practice.'* Teacher 9, School 5.

This was also reflected in the survey responses where respondents indicated that the opportunity to share and reflect upon practice was one of the most valued aspects of the URLEY intervention (see **Table 31**). Some open-ended survey responses highlighted the benefit of comparing practices and ideas with other practitioners and using that reflection to inform their own implementation of URLEY. Time spent sharing and reflecting took place during the training days, between colleagues at the individual schools, and with the mentors during visits and distance mentoring. Similar to the case study visits, some practitioners highlighted that this valuable reflection time was often forgotten in an otherwise hectic teaching schedule but that URLEY had made them prioritise it. As well as sharing with teachers from other schools, interviewees highlighted that the training afforded the opportunity to discuss and reflect as a team within the school.

However, survey respondents and case study interviewees noted some negatives and provided suggestions for improvements to the training sessions. Some said the training days had been too long and too many, and that the course content could have been covered in a shorter space of time. Some were 'concerned' by the amount of time out of class, particularly when a supply teacher was covering:

> *'That's one week of children's learning … We got supply teachers in but it's not the same is it? We've lost one week of learning and I don't think it's gained one week. That's all three classes lost that much.'* Teacher 7, Case Study School 3.

Some survey respondents said that a whole day could be overwhelming and result in information overload. In particular, some commented that the use of the same videos felt unnecessarily repetitive:

> *'Watching the repeated video clips (some more than 10 times) was tiresome and unnecessary.'* Early Year Lead and Nursery Teacher, Survey respondent 110.

Given the amount of information to apply, some would have liked a more structured set of goals to achieve and mentor visits to follow up quickly after sessions to maintain momentum, especially as when teachers return to class 'something else shouts louder and takes priority' (Teacher 8, School 4). This lower engagement between sessions made it harder for some teachers to fully participate in the training sessions:

> *'It was more, "Share your ideas, talk about what you're already doing (or something) with other schools." But if you hadn't necessarily done anything [since] the last one, and because of various other things that come up, you didn't really have an awful lot to share*.' Teacher 10, School 5.

Researchers also observed that during the training, it was often the same teachers who spoke out, or were called upon to share their ideas.

Some survey respondents also said that the afternoon sessions had been less helpful and could have been more practical. Although many valued the reflection and discussion times, they also found them 'long' and there was 'too much time chatting'. It was also noted, in two separate observations, that a minority of attendees used the time to chat 'off-task or discuss unrelated issues from the school'. This perception that the training would be 'drawn out' affected some teachers' decision to attend:

> *'It could have been condensed … I think we turned up, flicked through [the materials] and then decided if we needed to leave early.'* Teacher 1, School 1.

It is unclear whether this is related, but during the observed sessions, the researchers did notice that a handful of teachers left after lunch at each session. The most popular recommendation from the case study teachers and survey was to make the training more condensed. However, notes from observations suggest that the attendees had requested longer reflection times, which suggests the delivery team may have received mixed messages, or signals from the most engaged as opposed to the majority.

In terms of recommendations to improve the training, teachers mentioned it would have been useful to involve teaching assistants in the training, with some suggesting whole-school training days or individual TA training. Some suggested alternating training visits between schools to see as many perspectives as possible. The idea was that being exposed to as many settings and classroom environments as possible could have aided the valuable process of sharing and reflection between practitioners. There was a suggestion to rotate the training—to hold it at different schools to facilitate this. Many also cited that they valued their mentor being present at the training and would have liked that more consistently. Others mentioned the benefit of the autumn refresher day and wondered if all teachers should attend that to review their strategy for the new academic year. In a similar vein, another respondent suggested a reunion training day to ensure that practitioners kept their focus on the programme.

Materials and resources

Other support mechanisms were the physical and online resources. Materials downloaded included an online version of the pack/folder as well as other materials including the DVD clips from the training day. Overall, **Table 33** shows that the online resources were considered useful, with two thirds of survey respondents saying they were 'quite useful' or 'very useful'. Around a third of survey respondents were 'neutral'.

Table 33: Overall, how useful did you find the online resources? (N = 56)

	Number and percentage
Very useful	13 (23%)
Quite useful	24(43%)
Neutral	18 (32%)
Not very useful	1 (2%)
Not useful at all	0 (0%)

Another survey question asked how often respondents had used the online resources. The results are shown in **Table 34**.

Table 34: How often did you use the online resources? (N = 56)

	Number and percentage
Very frequently (more than once a month)	1 (2%)
Quite frequently (once a month)	13 (23%)
Three to five times	19 (34%)
Once or twice	23 (41%)
Never	0 (0%)

The online metrics from A+ Education showed that the total number of documents downloaded varied from a minimum of 1 to a maximum of 80 per school. On average, the total number downloaded was 13 (based on the 53 schools for which this information was available). The mean average number of downloads per 53 schools came to 8 documents. Although we have no way of knowing how the downloads were used or distributed, it does suggest a significant variation between settings, and potentially between staff access to the materials.

Open-ended survey responses and case study visits elaborated on these findings. Teachers had used the materials for a variety of purposes. In particular, teachers said the online copies of URLEY documents were useful for cascading the information to other staff. This included using videos and other materials for internal training days and displaying Language Learning Principles (LLP) posters around the learning areas. Other purposes were to help with planning, using the rating scales booklets, recapping learning between and after training sessions, and watching video clips. Others said that they had not used the online copies very much or at all, either due to time constraints or simply because they preferred the physical folder provided during the training.

Practitioners said that while discussions about the research readings during the training days were useful to gain a deeper understanding of the theory behind language development, the research readings themselves had not been very useful: the day-to-day school life was too hectic to have time to read them. Some suggested reducing the amount of readings and materials, and, in particular, making the research readings more manageable and shorter in order to reduce the workload of teachers and make it easier to disseminate core messages from the research readings to teaching assistants.

Practitioners said the DVD clips helped to show how the theory looked in practice and to show how to use the rating scales. The videos were also used as a refresher and made cascading easier as they were easily available to share with other staff; they also functioned as a good basis for discussion and reflection about good practices. However, there were also a number of negative comments related to the DVD clips, especially that the content was quite repetitive and some of the clips were not of a high quality. Some teachers said that while it was useful to see someone's else's practice, no early years classroom was the same so it could be hard to transfer the insights to their own classroom.

Finally, survey respondents were also generally positive about the accessibility of the online resources. Around half of survey respondents were neutral, responding it was 'neither hard nor easy' to access the online resources, while the other half said it was 'easy'. In open-ended responses, some teachers said that it was useful to have the materials in different formats. They had used the physical resources mostly, but it was useful to have an electronic copy that could be adapted.

Fidelity

Engagement and attendance

As discussed earlier in this report, the SAP specified a set of criteria identifying the minimum level of engagement required in order for the intervention to be considered to be taking place. This draws principally on school and teacher engagement scores and attendance data. As the intervention is focused on the Early Years Foundation Stage (EYFS) phase, it considers attendance data from both nursery teachers and reception teachers as well as engagement data from the phase as a whole. The relevant questions were:

1. Did the pupil have a nursery teacher who attended more than three training sessions?

2. Did the pupil have a reception teacher who attended more than three training sessions?
3. Was the school compliant with the intervention? An engagement score of 1 or 2 was considered to indicate compliance; a score of 3 or 4 as non-compliance. This was assessed by the mentors (part of the intervention team) who worked closely with each intervention school.

Table 35 presents the teacher attendance data at the training sessions. On average, around 69% of teachers were regarded to have complied with the attendance requirements set out in the protocol.

Table 35: Attendance

Teacher/Attendance	Nursery	Reception
Attended 3+	42 (70%)	59 (68%)
Attended 0–2	18 (30%)	28 (32%)
Total	60	87
Total %	100%	100%

Engagement

The school's engagement with the URLEY mentoring is defined as follows:

1. good, consistent engagement;
2. reasonable engagement—or, where this has been mixed, half or more of participating teachers have engaged;
3. some engagement—or, where this has been mixed, fewer than half of participating teachers have engaged; and
4. little or no engagement.

The mentors ranked their schools' engagement individually in summer 2016, after the training and mentoring had been completed.

Table 36: Engagement scores

Overall engagement score	Number
1	15 (25%)
2	14 (23%)
3	24 (40%)
4	7 (12%)
Total	60

This suggests that just under half the schools (48%) engaged with the intervention as intended.

Implementation environment

Support from senior leadership

The survey and case study interviews explored whether teachers and EYFS leads felt supported by the SLT to implement URLEY. **Table 37** shows that 86% of respondents felt that the SLT had been either 'quite supportive' or 'very supportive'.

Table 37: How supportive were your head and senior leadership team of the URLEY programme and your work to implement it? (N = 56)

	Number	Percentage
Not at all	0	0%
Not very much	3	5%
A little	5	9%
Quite supportive	18	32%

Very supportive	30	54%

This was elaborated in open-ended survey responses and in case study interviews. Teaching and support staff were asked to explain in what ways their headteacher and senior leadership team had been most and least supportive. The extent to which the SLT supported teachers to implement URLEY varied significantly across settings. Some teachers reported that the senior leadership team had given an appropriate amount of time for teachers to attend the URLEY training and for mentoring visits, as well as giving time for teachers within the setting to train other staff and to observe each other's practices. On the other hand, others reported that they would have needed more time outside class to appropriately reflect on the project, and that it was difficult to find time to arrange meetings to work on the project as a team. In one school, this was compounded by a reluctance to provide cover and release staff to have meetings about the implementation of URLEY, and, in some cases, to attend training.

Another theme was whether senior leaders had given the necessary financial support to make the necessary changes. Costly changes included adaptations to classroom provision identified by the environment rating scales or by the mentor, reported in the cost survey to extend up to £2,000. Some survey respondents specifically said that their SLT had made more funding available to make these changes. One case study respondent noted that her senior leader had been supportive whenever she had put in a request, but also noted that this had been facilitated by her own position as the EYFS lead in the setting. But there were also examples of teachers saying they had not been able to implement the necessary environment changes due to financial constraints, and in some cases teachers said that they would never even consider asking for costly environmental adaptations because they knew SLT would not be able to accept the request.

More generally, the issue of SLT financial support to implement any necessary changes was a recurring theme. Most responses were positive—that senior leaders had given the relevant team the freedom to implement strategies and adapt the environment as well as supporting changes to planning. Some respondents commented that senior leaders had been very interested in the project and supportive of the broader idea and philosophy behind URLEY. However, others said their senior leadership team had not been very involved and were not very knowledgeable about the URLEY project; some remarked that senior leaders had not been unsupportive, but that the support was primarily passive, just letting the teachers get on with it. For instance:

> *'We basically took it upon ourselves to just run with it at the time … It was just if I needed anything I would go and ask, but otherwise we lead it ourselves.'* Teacher 8, School 4.

However, others said that the limited SLT involvement had been detrimental to the implementation of URLEY and it would have been particularly beneficial if URLEY's importance had been raised across the whole school by the SLT. In particular, some of the mentors—who described SLT support as 'varied' across the schools—felt that SLT support was crucial in making URLEY succeed, though in some schools this was mitigated by having very strong EYFS leads and teachers who were trusted by the SLT. For instance:

> *'The schools that I felt were most successful in implementing URLEY were the ones where the heads had a deep understanding of good early years practice or had a great trust in their early years leaders and would enable them to make changes.'* Mentor 3.

Generally, the headteacher and senior leadership team did not attend the training days unless they attended in another capacity such as an EYFS lead or a teacher. One respondent said that their headteacher had attended one of the training days and her subsequent enthusiasm about the programme had facilitated a strong support from the senior leadership team throughout the project. For this reason, another respondent recommended that someone from the senior leadership team should be required to attend a minimum of one training day or be involved in a twilight session for the senior leadership team.

Finally, there were comments regarding how senior leaders, more generally, were engaged in the EYFS and in language development. Some teachers in case study schools said that their senior leaders showed strong passion for improving language development and showed a high level of awareness and recognition that their cohorts came in with really low levels of language ability. This recognition manifested itself in various ways; this included placing value on early years language development by signing up to interventions like URLEY, and for prioritising and allowing time for teachers to

implement the intervention. At the other end of the spectrum, the research team observed a few cases where teachers felt a negative pressure from senior management around numeracy and literacy.

Cascading the intervention to other staff

Typically, a couple of teachers from each school attended the training days and were then tasked with disseminating and cascading the URLEY programme to other teaching and support staff within their early years setting. The process evaluation explored how URLEY was cascaded through the survey and through interviews with teachers and support staff during the case study visits.

Table 38 shows that three quarters of survey respondents said they had disseminated URLEY 'quite a bit' or 'a lot'.

Table 38: How much have you disseminated URLEY to your wider team (for example, teaching assistants), for example by sharing information and materials, or involving them in implementing aspects of URLEY? (N = 56)

	Number (%)
A lot	14 (25%)
Quite a bit	28 (50%)
A little bit	11 (20%)
Not very much	2(4%)
Not at all	1 (2%)

The evaluation survey also asked respondents about the proportion of staff members who were actively engaged in URLEY. For schools with several responses from individual staff we used the average responses to compute the proportion of active URLEY EYFS staff in the 39 schools that answered the survey (this amounts to 65% of all treatment schools). **Table 39** shows that 44% of schools said all staff were actively engaged, and in 87% of settings, either half or more of EYFS staff were actively engaged in URLEY. Of course, this crucially depends on respondents' interpretation of what it means to be 'actively engaged in URLEY', but it does provide some indication as to how URLEY was felt to be implemented across the settings.

Table 39: Proportion of staff (teaching and support staff) within EYFS who are actively engaged in URLEY (N = 39 schools)

	Number and percentage of schools
0–25%	2 (5%)
25–50%	3 (8%)
50–75%	12 (31%)
75–100%	5 (13%)
100%	17 (44%)

Table 40 shows to what extent survey respondents found it challenging or easy to cascade URLEY to colleagues. This shows a mixed picture with relatively few respondents in the extreme answer options ('very challenging' and 'very easy') but instead 82% responded either 'quite challenging', 'OK' or 'quite easy'.

Table 40: How did you find cascading your learning from URLEY to colleagues (N=56)

	Number and percentage
Very easy	6 (11%)
Quite easy	16 (29%)
OK	17 (30%)
Quite challenging	13 (23%)
Very challenging	3 (5%)
I didn't try	1 (2%)

Through open-ended survey responses and during the case study visits, the evaluation explored in what ways teachers had found it difficult or easy to cascade URLEY.

A number of teachers explained how the programme design had helped them with cascading. Some teachers said that the materials and resources helped dissemination by making it easy to share the resource pack with other teaching and support staff and to put up displays in classrooms and staff rooms. Another key dissemination mechanism had been through arranging specific sessions or by having URLEY on the agenda during team meetings where everyone was told about the URLEY programme and its principles. Some of the teachers said the URLEY training had made them confident in disseminating the information effectively. Sometimes, the cascading had been in more informal settings such as during planning meetings or lunch breaks. Another main dissemination mechanism was through the mentors supporting the cascading process during visits—particularly by helping teaching assistants (TAs). This involved modelling and observation as well as helping with identifying improvements to classroom provision such as the home corner.

The key barrier to effective cascading was time constraints. Respondents said it was difficult to fit in conversations about URLEY in a busy school day and that it was difficult to get everyone together at the same time. This meant that some schools had not met, as a team, to discuss the implementation of URLEY. Teachers in these schools recognised that this would have benefited them as a team. Some of those who had arranged team meetings said that they still had not had enough time. Teachers said it was particularly difficult to cascade the intervention to TAs as they primarily work part-time and are only paid for their contact hours, and any additional time was spent doing lunchtime supervision or after-school clubs. In some schools, support staff did not attend staff meetings. This meant that the URLEY principles were not passed on properly to TAs:

> *'At the moment we've not got time to pass that knowledge on. We're struggling as it is passing day-to-day stuff on.'* Teacher 2, School 3.

One mentor said a couple of headteachers had paid the TAs an extra hour every six weeks to attend meetings, which had worked well. An EYFS lead in a case study school said their TAs went 'above and beyond' by coming in early, leaving late, and coming to the school during holidays, which meant there had been time for cascading.

The difficulties with finding time to cascade the URLEY information to TAs meant that some teachers resorted to simply modelling the principles during class hours. They recognised that this was not the most effective approach and emphasised that it was due to time constraints:

> *'It is hard, but I am trying to be a good role model; that's all I can do at the moment.'* EYFS lead and teacher, School 2.

Some of these teachers also relied on the mentor to disseminate the URLEY principles to teaching assistants during the visits, as described above. The challenges in cascading the URLEY intervention to teaching assistants were apparent when the research team visited case study schools. Often, teaching assistants didn't have much, or any, knowledge of the URLEY programme. This was confirmed by the mentors who said that it varied significantly to what extent the URLEY principles were cascaded to teaching assistants. Generally, however, it was fairly difficult to assess the teaching assistants' knowledge of URLEY. Some teachers said that they had not cascaded the principles with reference to the URLEY intervention but more generally as good language practices and sometimes in conjunction with other similar programmes. This meant that many teaching assistants mostly associated URLEY with the mentoring visit even in cases where teachers said they had cascaded the principles.

Another challenge was the buy-in and willingness to change practices among colleagues who had not attended the training. While some teachers said that all staff had been on board, others said that some staff members had been reluctant to change their practices—particularly that some experienced teaching assistants were 'stuck in their ways'. Another key challenge was to change TAs' language when this was not grammatically correct. Teachers found this hard to broach in a tactful way:

> *'The main issue we had was the incorrect grammar used by support staff. Challenging this was difficult and is a tricky thing to re-enforce positively.'* Reception Teacher, Survey respondent 59.

> *'It is really hard if you've got someone who speaks a certain way naturally; you know, it's kind of rude to say to someone that you can't say that. So, it was hard to get that balance right to be honest.'* Teacher 10, School 5.

One case study school said the URLEY programme had helped them with this difficult task:

> *'URLEY definitely gave the team … more confidence to kind of say to each other if they were perhaps not using a language that was grammatically correct … We were more confident to say, "Remember URLEY", and we turned it into a bit of a laugh, a bit of a joke.'* Teacher 9, School 5.

Some teachers suggested that some of these challenges could have been mitigated by sending more team members on the initial training course, or rotating staff members between training days. A few teachers also said that more support from the senior leadership team could have mitigated some of these difficulties, especially by allocating and prioritising time for dissemination.

Rating scales and TROLL

One of the main components of the URLEY intervention was to introduce the environment rating scales (ECERS and SSTEW) into the early year settings. **Table 10** showed that the vast majority of survey respondents said that the ratings scales had been useful in supporting and improving their practices. The open-ended survey responses elaborated that the rating scales had helped practitioners to evaluate and reflect on their current practices and environment, had helped to identify gaps, and helped to enhance their classroom provision. Some respondents noted that they still used them after the programme had officially finished while some respondents noted that it was difficult to do this.

We also explored the experiences of using the rating scales during the case study visits. Similar to survey respondents, many said the rating scales had been a very useful tool which had helped them evaluate and inform their practices. Examples of changes to practices due to the rating scales included making snack time more valuable for pupils, having more open-ended resources, improving the environment such as the home corner, and, more generally, helping to put the Language Learning Principles (LLPs) into practice.

However, the case study visits also showed that the knowledge and use of the rating scales varied significantly across schools. This was confirmed by mentors who said that some schools had hardly used the rating scales while other schools had used them effectively to monitor and inform their practice. At the very least, mentors used them alongside staff during mentor visits, so all schools used some of the rating tools at some point during the intervention. Often, practitioners said that it had been really useful to do observations alongside mentors, or having the mentor observe them, but sometimes this seemed to be their only experience of using the rating scales.

Most often, practitioners cited time as a barrier to using the rating scales as it requires a formal, three-hour observation. Settings had sometimes mitigated this by focusing on specific items on the rating scales that were particularly important to the setting such as those items related to the physical environment or the LLPs. This selection strategy had been recommended by mentors, which seemed like an acceptable adaption. To save time, one case study setting had gone through the rating scale booklets through discussion between practitioners without the observation. Another barrier seemed to be that some practitioners perceived it as an overwhelming exercise and some expressed insecurity about whether they used the rating scales properly. However, most seemed to emphasise that the training sessions had many practice sessions that had made them confident in using them.

The evaluation team visited case study schools the autumn term after the intervention had formally finished (2018). At this point, few schools still actively used the rating scales. Some teachers argued that their use of the rating scales during the intervention had already identified potential improvements and they did not feel they needed to use the scales again yet. An EYFS lead said:

> *'I think we are now URLEY'ed and we probably do it without even thinking about it. If I was to go into a setting where language was poor, staff weren't on board, environment was rubbish, then I would get the ECERS and the SSTEW out.'* EYFS lead, School 1.

Other teachers said it would be useful, particularly when introducing new staff to the provision, but sometimes pointed out that it was hard to fit into a hectic school day.

Similarly, the tool to assess pupil ability and progress (TROLL) was well-received among survey and case study respondents. Teachers felt it served as a useful and effective starting point to identify 'invisible' children and said it could be used again and again with new cohorts. However, some expressed concerns about finding the time to use the tools. A typical comment was:

> *'Doing a TROLL assessment has been useful as an initial guide to flag any children who may need more support, however finding the time to repeat the process is difficult.'* Reception Teacher, Survey respondent 37.

Among schools that were less engaged in the environment rating scales, practitioners often seemed to conflate the rating scales and the TROLL assessment, though this could also have been a result of the time that had passed since the intervention.

Language learning principles

The Language Learning Principles (LLPs) were frequently described as 'very useful' and 'invaluable' by survey and case study teachers. Many said that they used them regularly and they had become embedded into their practice. Practitioners said the LLPs had enabled them to reflect on their teaching practices, helped to inform their classroom practices, helped them to focus on encouraging different types of talk, and to ensure pupil language development.

Many schools had put posters and displays up in classrooms and staff rooms to remind practitioners of the importance of using the LLPs. The most engaged schools seemed to have integrated the LLPs into their planning, using the LLP terms and the logos throughout the planning process and in their planning documents.

Many teachers highlighted one or more LLPs that had substantially improved their teaching practice, and it was not uncommon for teachers to say they would never teach in the same way again. Some of the most frequently mentioned were:

- being a magnet for communication and being a 'language radiator' who modelled language for the children—including by extending children's language and by doing running commentary, and, more generally, using good language structures and vocabulary;
- talking at a level just above children's current level;
- providing repeated opportunities for pupils to bump into, and use, new words; and
- creating irresistible and meaningful contexts.

In addition, practitioners mentioned various conversation strategies such as 'observing, waiting, and listening' (OWL) and using commentary rather than questioning. Examples of teachers highlighting specific takeaways from the LLPs are:

> *'My biggest one is the commentary during children's play … Watching what they're doing, going into their play and then commentating on those children who don't have the language. That's the biggest thing that I found was the most useful thing to do.'* Teacher 1, School 1.

> *'The one that really stood out for me is the magnet for communication, just making sure that it's always there and ready for opportunities. But also making sure that the children do get access to new language.'* Teacher 10, School 5.

> *'OWL has been invaluable to me; I consciously make the effort to observe a situation and wait and listen before I intervene. This has at times shown me that the children can resolve situations by themselves when given the time to do so.'* Reception Teacher, Survey respondent 59.

The balance between LLPs and rating scales

Our case study visits showed that, for some participants, the LLPs were seen as the main component of the URLEY intervention. While the rating scales were generally considered useful, our case study findings suggest that the key takeaway for many participants were the language learning principles. The mentors confirmed this finding. They said that often the principles had been at the forefront in the mindset of practitioners, and that the environment rating scales had maybe been used less than anticipated. Mentors explained that the LLPs had really struck a chord with practitioners who especially liked that fact that LLPs provided clear guidance as to how they could improve their practice. However, mentors also emphasised that the connection between the rating scales and the LLPs had been effective:

> *'The Environment Rating Scales help us to shine a light on how well we are implementing those principles.'* Mentor C.

In the survey, teachers were asked an initial open-ended question exploring what they saw as the three key messages from the URLEY programme. Again, the responses showed how well-received the LLPs had been, but broadly, the responses suggest that practitioners understood the aims of the programme including the role of the Environment Rating Scales.

The main themes in the responses were URLEY's emphasis on the importance of language development. Most responses focused on how the LLPs emphasised the importance of a language-rich environment, communication between children and between adults and children, and the importance of bumping into new words and new language regularly. Respondents frequently referred to research findings that children's early language development has long-term impacts on their progress in school and their subsequent life outcomes.

As such, one of the key takeaways was the importance of providing a language-rich environment to enhance children's language development. In particular, the importance of interaction was recognised—both in regard to child-to-child communication where practitioners should 'be a facilitator for children to talk together' and, especially, that high-quality adult-child communication was key to a child's early language development. Respondents said that URLEY had instilled in practitioners the importance of prioritising more frequent and longer interactions with children. Many responses focused on how higher quality adult-child communication could be achieved, such as being a 'magnet for communication' by being a responsive language partner. The importance of 'observing, waiting, and listening' was frequently mentioned with reference to the importance of allowing time for the child to answer.

In addition, responses focused on the implementation of various strategies to achieve this environment. These were mostly focused on the LLPs such as being a 'language radiator', a 'magnet for communication', and various conversational strategies, and slightly less on the resources that could assist the formative assessment of pupils' individual language development and classroom provision such as the rating scales and the TROLL. The latter response focused on the importance of tracking and monitoring the progress of children and the classroom provision using tools such as the rating scales and TROLL, and generally tuning into children and each individual child. This helped identifying reluctant communicators and 'invisible' children, and participants emphasised the importance of the language development strategies for this group.

Embedding the URLEY programme into school practice

The case study visits were carried out after the intervention had formally ended. Practitioners were asked whether they were still implementing the URLEY intervention and whether they planned to do so in the future. Most practitioners said that parts of the intervention—particularly the LLPs—were still being implemented. Many treatment schools remarked that their cohorts typically had poor language ability which meant that language development remained one of their highest priorities. As such, the URLEY principles were integral to their everyday work:

> *'Once you've been on a course like that I don't think you can forget it completely to be honest. Language development underpins everything, doesn't it? We'll definitely still be using it.'* EYFS lead and teacher, School 2.

Some schools, especially those that seemed particularly engaged in the intervention, said they were no longer actively implementing the URLEY programme, but insisted the URLEY approach had been embedded to the extent that practitioners were using it subconsciously:

> *'I think it was something that we did and at the time we did it, we found it really beneficial. But we do, without thinking about it, still find it beneficial. It is something that we will always probably use, because you just do it without thinking now.'* EYFS lead and teacher, School 1.

Other schools said they were considering repeating the use of the rating scales, the TROLL assessments, and the conversation audit. Some said this was particularly important due to high staff turnover. Some of the mentors remarked that high staff turnover was the greatest barrier to whether the intervention would have a long-lasting legacy in the individual settings.

Outcomes

Perceived impact on pupils

The process evaluation also explored the perceived impact of URLEY on various outcomes, including on children, teachers, support staff, and the classroom environment and provision. Regarding the impact on pupils, **Table 41** shows that survey respondents thought that URLEY had had a direct impact on children, with 88% answering that it had 'quite a large positive impact' or a 'very large positive impact'.

Table 41: To what extent has the URLEY programme had a direct impact on the children in your class? (N = 56)

	Number (%)
Very large positive impact	15 (27%)
Quite a large positive impact	34 (61%)
Small positive impact	4 (7%)
No change	3 (5%)
Negative impact	0 (0%)

Through an open-ended question, the survey explored what types of impact on pupils had been observed. There was a large variety of responses. Some responses said that the intervention had most importantly led to a change in practices among teachers, which had the potential to improve pupil outcomes. Some examples of improved teaching practices were better support for pupils' language learning, an increased awareness of the classroom environment, using various techniques in their interaction with children such as the OWL ('observe, wait, listen) technique, thinking out loud, and using running commentary rather than questioning, and some said they had an increased focus on vocabulary including using vocabulary-enhanced texts, new vocabulary cards, promoting reading time, and providing irresistible contexts and opportunities to bump into new words. Some respondents also focused on how they had improved their formative assessment techniques by becoming more aware of the ability of their children and able to identify 'invisible' and quiet children. This included using the TROLL tool. Generally, respondents said URLEY had led teachers to focus more on children's language development. A typical comment was:

> *'URLEY has put language in all areas of the curriculum and at the heart of all our teaching and everything we do.'* Reception Teacher, Survey respondent 120.

Specifically, some respondents said that teachers had become more aware of interactions and that they had changed their planning and environment to ensure a more language-rich environment with more frequent interactions and better-quality interactions. This included more adult-child interactions in which staff listened more carefully to children by 'tuning in', using various questioning strategies, and generally creating 'more of a talking climate'. A few respondents said that this had led pupils to becoming better and more confident communicators, specifically those who were 'reluctant' or quiet:

> *'I had six children that I really focused on, you know the ones that just get on and do and that was good because they were more confident. You could see them talking more to their friends, coming to touch me, coming to show me things in the morning. I think that made a big difference to that small group of children.'* Teacher 10, School 5.

Generally, language confidence among pupils was suggested as one of the observed improvements. Survey respondents said some children had especially improved their confidence in initiating conversations and as a result their ability to form friendship groups. Other respondents focused on pupils' improved language, vocabulary, and writing skills, and that they had developed a love for the language in which they loved learning new words and using it in play.

> *'The children who started in nursery during the URLEY program implementation have developed a much broader vocabulary and a love for language that the other children don't have. They love learning new words, finding out the meaning, and using this in their play.'* Nursery Teacher, Survey respondent 10

A number of respondents said the impacts—such as improvements in confidence, vocabulary, and the ability to initiate a conversation—had been observed more among certain groups of learners; in particular, among EAL and language-deprived children, lower-ability pupils, reluctant speakers, and 'invisible' children:

> *'My lower ability now access the different areas and the communication is progressing and it's not so much the one word, but it's that, "Look what I've done!" It's getting bigger and bigger as we go on. There is definitely progress. Not for all children: I feel for my perhaps higher ability, maybe it's too much for them at times, but my lower ability definitely.*' Teacher 3, School 2.

Perceived impact on teachers and teaching assistants

The evaluation survey also asked questions more directly about the impact on teachers themselves. **Table 42** shows that 88% of respondents thought the URLEY intervention had 'quite a large positive impact' or a 'very large positive impacts' on them as a teacher.

> *'URLEY is a wonderful project in that it is so clearly defined what its best practice is. It is really rooted, I believe in best practice that in some ways that confidence that it gives to the teachers about why they're doing it ... with expertise brings confidence, with confidence brings a very strong early years*.' Headteacher, School 3.

Table 42: Overall, what impact do you think the intervention had on you as a teacher? (N = 56)

	Number (%)
Very large positive impact	16 (29%)
Quite a large positive impact	33 (59%)
Small positive impact	6 (11%)
No change	1 (2%)
Negative impact	0 (0%)

Survey respondents were prompted with a number of statements about to what extent URLEY had changed them as practitioners, and asked to rate to what extent they agreed with each statement. Statements with more than 75% agreeing were:

- 'The programme has made me feel more motivated about teaching': 45 respondents out of 56 (81%).
- 'The programme has given me a clearer vision of what I want to achieve as a practitioner': 51 (91%).
- 'The programme has made me feel more confident as a practitioner': 48 (86%).
- 'My understanding of how children learn and develop has improved/deepened': 50 (89%).
- 'My practice has improved because of the URLEY programme': 51 (91%).
- 'The programme has made me more open to changing my practice': 46 (82%).
- 'I am more successful in supporting children's learning': 47 (84%).
- 'I am more reflective or analytical about my practice', 49 (88%).
- 'I am more able to act on my reflections and put changes to my own practice into action': 48 (86%).
- 'I am more able to lead changes to pedagogy and practice within my class and/or within the EYFS': 45 (81%).

URLEY bought a consciousness of practice to teachers and TAs, particularly around their own 'personal interaction and practice' with children:

> *'I think it just makes us more conscious of the language that we use, extending it to words that they act and play, they might not hear and at their age they might not have heard of and try and enhance vocabulary and communication with the children, even though they speak to themselves.'* Teaching Assistant 1, School 1.

Raising awareness of why to do something made a particular impression. Both teachers and TAs explained they often did something instinctively, but the URLEY programme made them better understand why:

> *'I'd say it's improved in the sense of understanding the vocab and why the children need to hear it. Things like not correcting them, which we don't do anyway ... because we've done so much on language since I've been there, it's been a focus. We already know if they say I "like-d that", don't correct them, but say "I liked it". We try and get it embedded really. We've had such a change of staff as well it would be a hard one to answer. I'd say, yes, my practice has probably been more aware of how we speak, and things that are important ... You are made more aware of it, yeah. I'd say yes on the whole*.' Teacher 5, School 3.

URLEY was described by some as 'enlightening' and others as a 'useful reminder' of good practice with valuable 'hints and tricks'. In addition, two questions about whether teachers had experienced an increase in collaboration with colleagues inside and outside the classroom received responses with slightly less agreement, but still with more than 55% agreeing, and with most other respondents remaining neutral:

- 'I collaborate more often/effectively with colleagues within my class': 39 respondents out of 56 (69%).
- 'I collaborate more often/effectively with educators outside my class': 32 (57%).

Finally, one question asked about whether the respondent's sense of job satisfaction had improved, to which 60% of respondents agreed, with everyone else but three respondents remaining neutral.

Interviewees described that the URLEY programme made them more able to hold their TAs to account, particularly around their grammar and area specific language. Moving TAs away from questioning to commentary would also improve the interaction and experience of the children:

> *'We did a lot on role play and communication … we are taking more time to talk to them and to try and develop the language whereas before, well, we've always taken time but I'd say it's made us think a lot more about it. It's made us think a lot more and remind us.'* Teaching Assistant 4, School 3.

However, in some schools it was not a priority to cascade the intervention to TAs, which meant that the TAs seemed to progress less quickly that those who had attended the training.

Impact on classroom provision/environment

The evaluation survey also asked about the impact on the environment and classroom provision in terms of supporting children's personal and social development, their language and communication, and their literacy development. **Table 43** indicates perceived positive impacts, especially in terms of improving classroom provision to support language and communication.

Table 43: To what extent has the URLEY professional development had an impact on your classroom provision? Please rate each statement (N = 56)

	Support for children's personal, social, emotional development	Support for language and communication	Support for literacy development
	Number and percentage	Number and percentage	Number and percentage
Very large positive impact	16 (29%)	26 (46%)	12 (21%)
Quite a large positive impact	30 (54%)	25 (45%)	36 (64%)
Small positive impact	7 (13%)	4 (7%)	6 (11%)
No change	3 (5%)	1 (2%)	2 (4%)
Negative impact	0 (0%)	0 (0%)	0 (0%)

From the case studies, mentor interviews, and cost benefit survey, significant changes were made to some classrooms—with some buying additional items for up to £2000.

> *'We think construction needs to be moved away from the role play area. Let's put writing nearer the role play. So, we made changes within our setting because of the URLEY input.'* Teacher 9, School 5.

Teachers and teaching assistants found that changing the physical space evolved the way children interacted and communicated with staff and each other. Mentors were cited as particularly valuable to both make these changes and help upskill staff in how to use these areas with the children. Additionally, the posters and prompts were used as aids for both staff and children to support language and communication.

Existing practices: Was it business as usual?

The survey asked both treatment and control schools about how their approach to pupil language learning was similar/different to the one they had before September 2016, at the start of the URLEY intervention. We report on control group activity later in this section. Among treatment schools the responses were fairly mixed, with relatively few responses at the extreme options ('very similar' and 'very different'), and slightly more responses which highlighted that practices had changed over this period.

Table 44: To what extent is your approach to pupil language learning similar/different to the one you had before September 2016? (N = 56)

	Number
Very similar	4 (7%)
Quite similar	15 (27%)
Quite different	28 (50%)
Very different	9 (16%)

This was elaborated during interviews with case study schools and mentors. Most schools had not used the environment rating scales prior to the intervention, and many had not heard about them. For the few schools who had used the rating scales, typically this had involved the visit of an early years specialist from the Local Authority or from a neighbouring school that had assessed them on ECERS or SSTEW. And even this had been fairly limited and was usually a one-off occurrence. A mentor commented:

> *'I think some had used them before, but not in this very focused way to draw out the particular items that will be strong for language development.'* Mentor 3.

In terms of the language principles, some of the mentors said that the approaches to language development were quite varied across the schools. It was common for teachers in case study schools to say that many of the LLPs were not revolutionary or ground-breaking, but it had made them more mindful practitioners, and made them tweak or reinforce good practices. Often, teachers identified large improvements in very specific areas such as improving their questioning strategies and commentating on children. Some teachers also said that it had been beneficial to improve their understanding of the theory and pedagogy behind good practices which had helped them to become more intentional in their teaching practices. Mentors also said that it had made teachers more aware of particular strands of language development. Often, practitioners were thinking about language as a whole rather than particular strands.

Finally, one of the mentors had observed that at the start of the intervention some practitioners associated observations with negative experiences that were used as a stick to beat them with. These schools had experienced a culture and mind shift by seeing how the rating scales used observation as a positive and formative tool.

Overall experiences of URLEY and recommendation

The final question of the evaluation survey asked respondents whether they would recommend the URLEY programme to others. **Table 45** shows that, for the initial sample, 91% of respondents answered 'probably yes' or 'definitely yes'. This is based on the responses from 56 teachers out of 144 (39% response rate). After closing the initial evaluation survey, we followed up all non-respondents and asked them this question, and received 26 further responses, taking the total up to 82 respondents out of 144 (57% response rate). The follow-up responses showed the same pattern, with a large majority recommending the intervention.

Table 45: Would you recommend the URLEY programme to others? (initial sample: N = 56 and follow-up: N = 82)

	Initial sample	Initial sample + follow-up
	Number and percentage	Number and percentage
Definitely not	0 (0%)	0 (0%)
Probably not	0 (0%)	0 (0%)
Unsure	5 (9%)	6 (7%)
Probably yes	14 (25%)	19 (23%)
Definitely yes	37 (66%)	57 (70%)
Total	56 (100%)	82 (100%)

Formative findings

The findings highlight a number of barriers, and potential ways these could be addressed.

The main barrier was that the URLEY programme required a significant time investment. In particular, attendance at training and engagement with the intervention was time-intensive and difficult to sustain for a number of teachers. This could be addressed by making training days shorter and fewer, and generally condensing the training, for example by allowing less time for reflection and focusing more on practical activities. However, this would of course have implications, including having less time for reflecting on practices and understanding the theory behind the URLEY intervention.

Another identified barrier was the challenge of effectively cascading the intervention to staff who did not attend the training, including teaching assistants. This was described as challenging unless the SLT had provided timetabled space for this to take place. Additional training for these staff may have helped improve their knowledge about the programme and their engagement. Similarly, future implementations of URLEY should consider how to give practitioners more time to appropriately reflect on, and have conversations about, URLEY outside of classes, including as a team, as it was difficult to fit this into the normal school day. In addition, some practitioners said research readings should be made shorter and more accessible to improve the chances of fitting them into a hectic school day and to effectively cascade them to other staff.

Another issue was financial cost. Some schools had identified necessary environmental changes, but had not been able to implement these due to financial constraints.

Finally, the findings show that the LLPs were especially valued by practitioners, perhaps more so than the rating scales, but it was difficult to disentangle the exact relative importance of LLPs and the rating scales. Subsequent URLEY interventions should consider in further detail the role of the rating scales, including how frequently these should be used, how they best work with the LLPs, and how to address the fact that they are time-consuming for practitioners to implement.

Control group activity

Control survey

A brief survey was sent out to control schools at the same time as the research assistants conducted the language post-tests. The questions in the survey were designed to understand how the nursery environment of control school settings differed from that of the intervention group.

In total, 37 out of 60 control schools returned the survey, which represents over 60% of the total control group.

Did the control schools invest in early years language programmes during the trial?

Most survey respondents confirmed that they had. Three quarters (75%, 28) of schools that responded to the survey stated they had used resources to improve pupil language.

Seventy percent (26 schools) referenced particular Early Years Literacy programmes they had invested in. The most common was WellComm, with 54% (14) of those who cited a programme referencing that they used this—either standalone, or as part of a suite of options. The remaining schools cited Talk Boost (6), ELKLAN (4), and ERS (4). The programmes mentioned by one school each were Talking Tables, ICAN, Helicopter Stories, Pyjama Drama, Better Talking Partners, and EY2P.

How much has their approach to pupil language learning changed since 2016?

Twenty-seven schools (72%) said their approach was either 'quite similar' or 'very similar'. Given that URLEY aimed to embed improved practices in the setting, this suggests that the classroom environment remained static, and the majority of control schools did not instigate significant changes in this respect, despite many investing in other language programmes.

Among the 28% of schools that stated their approach was 'quite' or 'very' different, the reasons given were as follows: new targeted intervention programmes, staff training and a change to pedagogic methods, raising the profile of the importance of language in the home learning environment, and a new whole-school focus or intervention on language that filtered across all year groups. These approaches have the potential to be quite effective.

Did any control schools use the intervention materials?

Four of the control schools referenced using the materials from the URLEY programme, notably SSTEW and ERS; 34 schools responded to the question as to whether they had used ECERS or received support to use the ERS in the previous few years. Most of the latter—70%, 24 schools—confirmed they had not at all or had not used it very much at all. Just over a quarter—26%, 9 schools—said they had a little, and one school reported they had used it quite a lot. Those which had used the ERS cited using it for multiple purposes: auditing the setting or children (5), planning (6), or had received guidance or training on it (4). In any case, these cases are unlikely to contaminate the RCT design as the control schools' use of materials from the URLEY programme are unlikely to be as intensive as the URLEY intervention or supported by the same amount of training or mentoring.

How much do control schools know about the intervention?

Of the 36 schools that responded to the question, 'How much do you know about ERS? For example, ECERS, SSTEW?', the majority—63%, 23 schools—said they knew 'a little, not very much, or nothing' about ERS, ECERS or SSTEW. The remaining 13 schools felt they knew quite a lot, with one stating it knew a lot about the ERS.

Conclusion

Key conclusions
1. Children in schools receiving URLEY did not make additional progress in language development compared to children in control schools, as measured by a composite language score. This finding has a moderate to high security rating. The effect size is equivalent to one month's less progress than the control group, though is equivalent to zero months once imbalance on the numbers of FSM and EAL children in each arm is controlled for. The result was similar for pupils eligible for FSM.
2. The programme had a positive impact on quality of provision (as measured by Environment Rating Scales), with effect sizes in the range of 0.5 to 0.7. This suggests that quality of practice improved (for example, the quality of language-supporting adult-child interactions), but not at a sufficient level to translate to improved language outcomes for children. It may be that impacts on pupil outcomes would only be observed in the longer term, or with even larger improvements to practice.
3. Many children were not taught in reception by a teacher who had received the full training (partly due to substantial teacher turnover in the schools), and it was not possible to assess the extent and impact of this in the evaluation. Additional induction training was provided where possible, but this is nonetheless likely to have reduced the potential impact of the URLEY programme.
4. Teachers were overwhelmingly positive about the URLEY programme: 91% of responding teachers felt the intervention had a positive impact on the quality of provision and highlighted the mentoring as especially valuable. Many teachers felt the programme was most beneficial to a targeted subset of reluctant communicators, as opposed to whole-class improvements.
5. The URLEY programme required significant time investment and cascading the intervention to staff who did not attend the training was challenging. Condensed training and a more structured approach with milestones, goals, and senior leadership team (SLT) support may have helped teachers to prioritise the programme.

Interpretation

The current trial does not demonstrate evidence that the URLEY intervention improves children's language and social-behavioural outcomes.

The intervention was found to have no positive effect on the composite language outcome. The estimated effect size in the primary analysis is -0.08, indicating a small, negative, but non-statistically significant impact (the magnitude of the effect is equivalent to one month's less progress). This result was robust to sensitivity checks around missing data. Further analysis controlling for EAL pupils and whether pupils were eligible for FSM, in response to the observed imbalance in these characteristics, reduced the magnitude of the effect, indicating essentially no difference in progress between treatment and control groups (an effect size of -0.02). This robustness check suggests that the effect size observed in the primary analysis could have been driven by the imbalance in EAL and FSM pupils across treatment arms, rather than by the intervention itself. Analysis of the subgroup of FSM pupils indicated similar results among FSM pupils, with a small, negative, but not statistically significant effect (an effect size of -0.08).

Examining each of the component scales of the composite language score individually shows small, negative, but not statistically significant impacts on scores on the BPVS (effect size -0.058), the RAPT information (effect size -0.044) and RAPT grammar measures (effect size -0.049); these effect sizes correspond to one months' less learning. A negative impact of 0.122 (corresponding to two months' less learning) is observed on the CELF Sentence Structure test, however this is also not considered statistically significant.

In addition, we found no impact of statistical significance (equivalent to zero months' additional progress) of the intervention on social-behavioural development (effect size -0.021) as measured by the Adaptive Social Behaviour Inventory (ASBI) or its subscales.

The analysis did find that the programme had a statistically significant and positive impact on the composite Environment Rating Scale (ERS) measure, as well as its subscales. Previous research suggests that attending a higher-quality pre-school (as measured by ERS) is associated with increased attainment, both in primary and secondary school (Sylva et al., 2014). While this trial provides evidence that the intervention improves the quality of provision—at least as captured by the ERS—and could, therefore, potentially improve attainment outcomes and social outcomes in later years as per previous studies, it did not translate to improvements for pupils in the time-period considered in this study.

Given the nature of the intervention, which teaches practitioners to use research tools to evaluate the learning settings, it is perhaps not surprising that the intervention significantly improved the ERS ratings as teachers receiving the intervention likely became more mindful of the components of the scales in their classrooms. Furthermore, we found improvements across the whole ERS and SSTEW scales, which suggests that teachers may have deepened their underlying knowledge of child development and pedagogy.

Furthermore, this finding is consistent with the findings of the implementation and process evaluation, in which 91% of respondents felt the intervention had 'quite a positive' or a 'very positive' impact on the quality of provision (bearing in mind that non-respondents may not have shared the same view). In particular, practitioners felt the intervention made it easier to support language and communication skills. A key finding with regard to impact from the process evaluation is that URLEY, and particularly the implementation of Language Learning Principles, led teachers to improve their teaching practices, including their approaches to language learning and interactions with pupils. The teachers also perceived that this had fed into improvements for pupils, though obviously this has not been demonstrated in the impact evaluation. Perceptions of impact could relate to outcomes not captured by the measures used in this study, however, perceived impact is not the same as actual impact (for example, it would be difficult to attribute progress to the intervention rather than simply natural progress over time).

It is perhaps puzzling, therefore, that the observed improvements in the quality of provision (at least as assessed by the ERS) do not feed through to pupil outcomes in the impact analysis. Although we cannot definitively say why this is the case, it is useful to consider some possible explanations.

One potential contributing factor could be that treatment schools did not sufficiently comply and engage with the intervention as desired. Indeed, the process evaluation observed mixed levels of fidelity: around seven in ten teachers complied with the attendance requirements and just under half of schools (48%) engaged with the intervention as desired. Another related issue was that teachers who attended the training days found it challenging to cascade the information from training days to other staff members, particularly due to time constraints. This means that while the URLEY programme was highly valued by some teachers, it may not have been sufficiently implemented by all staff members to have the intended impact. However, our analysis of compliance does not indicate positive impacts on language development among those settings deemed to be compliant with the intervention.

Another possible explanation is that URLEY takes a longer time to embed and for impacts on children's language development to occur. During the case study visits, which took place after the formal intervention had ended, there was evidence to suggest that URLEY had been largely embedded in some schools, and in particular the Language Learning Principles were still informing teaching practices. The impact of environmental changes may also be delayed, particularly if they are implemented late during the programme: while the effect of the program may not be strong enough to improve pupil outcomes in the short term, it is possible benefits would be observed for future cohorts of children. Alternatively, it may be that individual children simply did not receive sufficient of the improved language-supporting interactions to make a difference to their development, for example because of the relatively low adult-child ratios in schools.

In addition, the programme was described as time-consuming by teachers, in particular the requirement to attend five whole training days. Therefore, while those teachers perceived the URLEY intervention to improve their teaching practices, these benefits may have been offset, at least in the short term, by reduced quality of teaching during hours out of class, and particularly during supply cover, which was indeed highlighted as a concern among some teachers. In the longer term, however, the higher teaching quality may feed into pupil outcomes for future cohorts. Any such effect may be tempered, however, by staff turnover, which was noted to be an issue within the trial and which may also have served to reduce the likelihood of observing an impact on children's outcomes.

Nevertheless, a positive impact on ERS scores is found, which previous studies have linked to improved attainment. In interpreting our results, it is worth bearing in mind that while we see a positive impact on the ERS measures, the average ERS scores among participating settings are not particularly high—a score of three on the ECERS, for example, would be considered 'minimal' and a score of five 'good'—and the average scores on the composite ERS measure (at follow-up) in this study stood at 3.5 for treatment settings and 2.9 for control settings. Perhaps impacts on outcomes would only be observed once settings achieved higher quality ratings. This, in part, helps to address why the intervention had positive and significant effects on ERS scores, but not on the chosen pupil outcomes.

In sum, while the effects on children's language and behavioural outcomes examined in this study were largely found to be null, the qualitative process evaluation resulted in overwhelmingly positive feedback around the process through which teachers became trained in the intervention and implemented elements of it into their classroom practice.

Limitations

This project faced slightly higher than expected levels of attrition at the student and school levels, in addition to a relatively high rate of teacher turnover in the intervention schools.

The project team had predicted 15% student-level attrition; however, by analysis stage, 22% of students had dropped from the sample. This was due in part to the five schools that dropped out of the study, which also reduced the number of clusters from the required 120 to 115. However, this attrition is unlikely to have influenced the balance between treatment and control groups at the analysis stage since we observed balance on all characteristics except percentage of pupils in FSM and percentage of pupils marked EAL at both randomisation and analysis. The treatment group had a higher percentage of EAL pupils than the control group. Subgroup analysis for EAL pupils showed lower overall scores on the composite language measure but did not indicate a differential effect of the treatment for this group. Once we account for the percentage of EAL and FSM pupils in the analysis, we see no difference in progress between the treatment and control groups.

While we conducted various robustness checks in relation to missing data, it is worth acknowledging that some children will be missing post-test data due to greater difficulties in being able to engage with the test. While this should not differ by trial arm, it may have some bearing on the results if it is children with the lowest levels of language ability who stand to gain most from the intervention. As a related point, some of those children missing pre-test data will also be those of lower language ability. This may mean that the school-level averages used as the pre-test measure could be biased upwards, although this would apply for both treatment and control groups.

Schools faced a high rate of teacher turnover, throughout the project but especially between the two intervention years, which is likely to have an impact on the extent to which the intervention could have an impact on children's outcomes. This is unlikely to change much in the control schools; however, if departing teachers in intervention schools were replaced with newer teachers who did not receive the same training, the effects of the intervention captured by the study are likely to be biased towards zero.

Future research and publications

Given the results around ERS scores, in spite of the null effects on the composite language measure and ASBI scores, future trials might consider measuring the impact of the intervention directly on various elements of teacher practice. As part of this, it may be worth considering whether there are particular elements of the programme that seem particularly effective. Exploring this might help to refine the intervention in such a way that it can more effectively help teachers receiving the programme to improve pupils' language and behavioural skills. The reported levels of teacher enthusiasm that were revealed in the process evaluation, but lack of strong positive effects on children's attainment, suggest that there is scope to harness this teacher enthusiasm to improve pupils' educational outcomes.

If studies like this were to be undertaken, it would be important to address the issues brought on by teacher turnover (which was significant in this study) and increase fidelity to the intervention design where possible. Specifically, this would mean to ensure that new teachers coming in to replace departing teachers in intervention schools are trained in the intervention materials at the same standard as the original cohort of participating teachers. It may be valuable to consider if there are ways in which the programme could be condensed in order to make participation more manageable for staff and to minimise time away from the setting (which could be informed by research into the most effective elements). The developer noted that many of these changes have since been implemented within the version of the programme currently being used within the DfE SSIF projects.

While the current trial did not identify impacts on language or behavioural development using the specified measures, it would be valuable to explore whether there were any impacts on children's performance as measured by the Early Years Foundation Stage Profile. In addition, and particularly given the positive impact on quality of provision, it would be of value to explore whether there are longer-term impacts on children's attainment.

It is the intention of the project and evaluation teams to seek to publish these findings. No further analysis by the evaluation team is anticipated.

References

Anderson, L. M., Shinn, C., Fullilove, M. T., Scrimshaw, S. C., Fielding, J. E., Normand, J., Carande-Kulis, V. G and Task Force on Community Preventive Services (2003) 'The Effectiveness of Early Childhood Development Programs: a Systematic Review', *American Journal of Preventive Medicine*, 24 (3), pp. 32–46.

Bierman, K. L., Domitrovich, C. E., Nix, R. L., Gest, S. D., Welsh, J. A., Greenberg, M. T., ... and Gill, S. (2008) 'Promoting Academic and Social-Emotional School Readiness: The Head Start REDI Program', *Child Development*, 79 (6), pp. 1802–1817.

Bierman, K. L., Nix, R. L., Heinrichs, B. S., Domitrovich, C. E., Gest, S. D., Welsh, J. A. and Gill, S. (2014). 'Effects of Head Start REDI on Children's Outcomes 1 Year Later in Different Kindergarten Contexts', *Child Development*, 85 (1), pp. 140–159.

Burger A. and Chong I. (2011) 'Receptive Vocabulary', in Goldstein S. and Naglieri J. A. (eds), *Encyclopedia of Child Behavior and Development*, Springer, Boston, MA.

Campbell, F., Conti, G., Heckman, J. J., Moon, S. H., Pinto, R., Pungello, E. and Pan, Y. (2014) 'Early Childhood Investments Substantially Boost Adult Health', *Science*, 343 (6178), pp. 1478–1485.

Cordingley, et al. (2015) 'Developing Great Teaching: Lessons from the International Reviews into Effective Professional Development'. https://tdtrust.org/wp-content/uploads/2015/10/DGT-Full-report.pdf

Currie, J. (2001) 'Early Childhood Education Programs', *Journal of Economic Perspectives*, 15 (2), pp. 213–238.

Dickinson, D., McCabe, A. and Sprague, K. (2003) 'Teacher Rating of Oral Language and Literacy (TROLL): A Research-Based Tool', *Reading Teacher*, 56 (6), pp. 554–564.

Dockrell, J., Llaurado, A., Hurry, J., Cowan, R., Flouri, E. and Dawson, A. (2017) 'Review of Assessment Measures in the Early Years', London: Education Endowment Foundation.

Domitrovich, C. E., Cortes, R. C. and Greenberg, M. T. (2007) 'Improving Young Children's Social and Emotional Competence: A Randomized Trial of the Preschool "PATHS" Curriculum', *Journal of Primary Prevention*, 28 (2), pp. 67–91.

Education Endowment Foundation (2018). 'Sutton Trust-EEF Teaching and Learning Toolkit' and 'EEF Early Years Toolkit'.

Gorey, K. M. (2001) Early Childhood Education: A Meta-Analytic Affirmation of the Short- and Long-Term Benefits of Educational Opportunity', *School Psychology Quarterly*, 16 (1), p. 9.

Heckman, J. J. (2012) 'The Case for Investing in Disadvantaged Young Children', European Expert Network on Economics of Education.

Heckman, J. J. and Masterov, D. V. (2007) 'The Productivity Argument for Investing in Young Children', *Applied Economic Perspectives and Policy*, 29 (3), pp. 446–493.

Hopkin R., Stokes, L. and Wilkinson, D. (2010) 'Quality, Outcomes and Costs in Early Years Education', Report to the Office for National Statistics.

Raver, C. C., Jones, S. M., Li-Grining, C., Zhai, F., Metzger, M. W. and Solomon, B. (2009) 'Targeting Children's Behavior Problems in Preschool Classrooms: A Cluster-Randomized Controlled Trial', *Journal of Consulting and Clinical Psychology*, 77 (2), p. 302.

Sibieta, L., Kotecha, M. and Skipp, A. (2016) 'Nuffield Early Language Intervention. Evaluation Report and Executive Summary', London: Education Endowment Foundation.

Springate, I., Atkinson, M., Straw, S., Lamont, E. and Grayson, H. (2008) *Narrowing the Gap in Outcomes: Early Years (0–5 Years)*, Slough: NFER.

Stoll, et al (2012) 'Great Professional Development Which Leads to Great Pedagogy: Nine Claims from Research'.

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/335707/Great-professional-development-which-leads-to-great-pedagogy-nine-claims-from-research.pdf

Sylva, K., Melhuish, E. C., Sammons, P., Siraj, I. and Taggart, B. (2008) 'Final Report from the Primary Phase: Pre-school, School and Family Influences on Children's Development During Key Stage 2 (7–11)', Nottingham: DCSF Research Report 61 / Institute of Education, University of London.

Sylva, K., Melhuish, E. C., Sammons, P., Siraj, I. and Taggart, B. with Smees, R., Toth, K. and Welcomme W. (2014) 'Effective Pre-school, Primary and Secondary Education 3–16 Project (EPPSE 3–16): Students' Educational and Developmental Outcomes at Age 16', Department for Education' Research Report RR354.

Timperley, et a.l (2007) 'Teacher Professional Learning and Development: Best Evidence Synthesis Iteration'. http://www.educationcounts.govt.nz/publications/series/2515/1534.

Webster-Stratton, C., Reid, M. J. and Stoolmiller, M. (2008) 'Preventing Conduct Problems and Improving School Readiness: Evaluation of the Incredible Years Teacher and Child Training Programs in High-Risk Schools', *Journal of Child Psychology and Psychiatry*, 49 (5), pp. 471–488.

Appendix A: EEF cost rating

Cost ratings are based on the approximate cost per pupil per year of implementing the intervention over three years. More information about the EEF's approach to cost evaluation can be found **here**. Cost ratings are awarded as follows:

Cost rating	Description
£ £ £ £ £	*Very low:* less than £80 per pupil per year.
£ £ £ £ £	*Low:* up to about £200 per pupil per year.
£ £ £ £ £	*Moderate:* up to about £700 per pupil per year.
£ £ £ £ £	*High:* up to £1,200 per pupil per year.
£ £ £ £ £	*Very high:* over £1,200 per pupil per year.

Appendix B: Security classification of trial findings

OUTCOME: ***Composite language measure based on: British Picture Vocabulary Scale, Renfrew Action Picture Test, Clinical Evaluation of Language Fundamentals (CELF) Preschool 2 UK***

Rating	Criteria for rating			Initial score		Adjust		Final score
	Design	MDES	Attrition					
5	Randomised design	<= 0.2	0-10%					
4	Design for comparison that considers some type of selection on unobservable characteristics (e.g. RDD, Diff-in-Diffs, Matched Diff-in-Diffs)	0.21 - 0.29	11-20%					
3	Design for comparison that considers selection on all relevant observable confounders (e.g. Matching or Regression Analysis with variables descriptive of the selection mechanism)	0.30 - 0.39	21-30%	3		Adjustment for threats to internal validity **[0]**		3
2	Design for comparison that considers selection only on some relevant confounders	0.40 - 0.49	31-40%					
1	Design for comparison that does not consider selection on any relevant confounders	0.50 - 0.59	41-50%					
0	No comparator	>=0.6	>50%					

Threats to validity	Threat to internal validity?	Comments
Threat 1: Confounding	**Moderate**	Randomisation was carried by an independent member of the independent evaluation team, and the relevant code is provided. Some evidence of imbalance in the pre-test are found (0.05 sd in the individual level, 0.11sd in the school level means), but these are accounted using school level means as the baseline for the primary outcome.
Threat 2: Concurrent Interventions	**Low**	No evidence of concurrent interventions but this is based on a relatively low response rates to questionnaire surveys
Threat 3: Experimental effects	**Low**	No evidence that participation in the study changed practice in control group schools
Threat 4: Implementation fidelity	**Low**	Well defined and reported. Evidence that not all schools and teachers completely implemented the intervention. Issues were mainly about lack of resources, lack of support from SMT and around cascading the approach to members of staff not attending the training
Threat 5: Missing Data	**Low**	Sufficient analysis is presented to suggest that different approaches to missing data do not result in different conclusions
Threat 6: Measurement of Outcomes	**Low**	Outcome measures seem well chosen for the target population. Assessments were carried out by trained research assistants blind to treatment/control allocation. However, a better explanation of the primary composite measure could have been included.

Threat 7: Selective reporting	**Low**	Trial was registered and protocol and SAP were followed with only minor (and documented) changes. Consort and TiDier requirements are clear, but a logic model would have been useful.

- **Initial padlock score:** 3 Padlocks – This was a randomised controlled trial with an MDES at randomisation of 0.22 and attrition at the pupil level of 22%
- **Reason for adjustment for threats to validity**: None. The main threat is around the imbalance in pre-test observed, but this is controlled for in the analysis and not substantial enough to grant a reduction in the security rating.
- **Final padlock score:** initial score adjusted for threats to validity = 3 Padlocks.

Appendix C: Recruitment materials

Memorandum of Understanding

Agreement to participate in the evaluation of
Using Research tools to Improve Language in the Early Years

Please sign both copies, retaining one and returning the second copy to [NAME OF CONTACT] at [PROJECT DELIVERY ADDRESS/EMAIL] by[ADD DATE]

School Name: _____

School Postcode: _____Headteacher Name: _____

Aims of the Evaluation

The aim of this project is to evaluate the impact of "Using research tools to improve language in the early years", a professional development programme designed to support teachers to assess and improve the quality of their practice. The results of this research will make an important contribution to understanding what works in improving language and social outcomes for children in the early years.

The Professional Development Programme

The professional development (PD) will be provided by the University of Oxford, University College London (UCL) and A+ Education Ltd, and is funded by the Education Endowment Foundation (EEF). It will provide nursery and reception teachers with specialist training in how to support language and social development in the early years. It will also prepare teachers to use research tools to evaluate practice within their classrooms, and use evidence-based strategies to develop aspects identified as needing improvement.

The PD comprises a five day course over two terms, with time between sessions to use the research tools in practice, and a single follow-up day in the third term. Sessions include examples of good practice, as well as materials and strategies to use in class. Individual support will be provided by a project mentor.

In each participating school, at least one nursery teacher and one reception teacher will receive the training (maximum three nursery/reception teachers per school).

Schools who agree to take part will be randomly allocated to either the Phase 1 (intervention group) or Phase 2 (comparison group) in the latter part of the autumn 2016 term.

- Teachers in Phase 1 schools will receive the training course in the spring and summer terms of 2017 (16-17 school year) and the follow-up training day in autumn 2017 (17-18 school year), with mentoring support provided throughout the year.
- Phase 2 schools will receive enough funding to access the programme from autumn 2018, or a payment of £1,000 if they wish to spend the funding on something else.

The Evaluation

The evaluation is being conducted by the Behavioural Insights Team (BIT) and the National Institute of Economic and Social Research (NIESR), collectively the 'research team'.

Random assignment of schools to Phase 1 and 2 is essential to the evaluation as it is the best way of outlining what effect the project has on children's outcomes. It allows the research team to compare progress made by children in Phase 1 and 2 schools before Phase 2 schools start the programme, to see what impact it has. It is important that schools understand and consent to the random allocation process. The key features of the evaluation are:

- The language and social development of all children in nursery classes in the participating schools (Phase 1 and Phase 2) will be assessed in autumn 2016. In summer 2018, the language and development of the same children (now in reception) will be assessed again.
- One class per school (Phase 1 and Phase 2) will be observed at the start of the project, in order to establish current practice, and again at the end of the project, to assess how this may have changed.
- All schools will also be asked to provide selected school and pupil level information in order to facilitate the evaluation and to enable linkage to the National Pupil Database.
- In addition, teachers (in both Phase 1 and Phase 2 schools) will be asked to complete a short questionnaire, and a small number of schools will receive visits from the research team to observe practice and conduct interviews with teachers.

Use of Data

All data, including pupils' test responses and any other pupil data, will be treated with the strictest confidence. Pupil assessments will be administered by BIT and accessed by BIT and NIESR. Named data will be matched with the National Pupil Database and shared with the University of Oxford, UCL and A+ Education Ltd, the Department for Education, EEF, EEF's data contractor FFT Education and in an anonymised form to the UK Data Archive. No individual school or pupil will be identified in any report arising from the research.

Responsibilities

The project team (University of Oxford, UCL and A+ Education Ltd) will:

- Deliver the five day course, plus one follow up training day, as well as providing individual support from a project mentor
- Be the first point of contact for any questions about the project
- Provide on-going support to the school

The research team (BIT and NIESR) will:

- Conduct the random allocation of schools to Phase 1 or Phase 2
- Collect and analyse the data from the project to estimate the impact of the intervention
- Ensure all staff carrying out assessments are trained and have received DBS clearance
- Publish a report on the findings of the project and disseminate research findings

The school will:

- Consent to randomised allocation and commit to the outcome, whether assigned to the Phase 1 (intervention) or Phase 2 (comparison) group
- Allow time for each assessment phase and liaise with the research team to find appropriate dates and times for assessments to take place
- Allow a class to be observed by staff from A+ Education Ltd, at both the start and end of the project.
- Release nursery and reception teachers so that they can attend the training sessions and access mentor support
- Ensure the shared understanding and support of all school staff for to the project and personnel involved

We commit to the evaluation of the Using Research Tools to Improve Language in the Early Years project as detailed above

Head teacher signature: _____ Date: _____

Information Sheet and Parental Consent Form

Children need good language skills to help them make the best of school and become confident and successful learners.
Your child's school is taking part in an exciting research study, designed to help teachers and other staff further develop their skills in supporting children's language and social development. This involves training teachers to use a set of rating scales used in research studies, to evaluate and develop their practice.

Nursery and reception class teachers will
receive specialist training and expert support from a mentor. This will be provided by a research team from the University of Oxford, University College London and A+ Education Ltd.
The study is being evaluated by the Behavioural Insights Team (BIT) and National Institute of Economic and Social Research (NIESR). The aim is to find out whether the programme improves quality and children's development. Knowing this will make an important contribution to understanding what works in improving children's language and social skills in the early years, to set them on the road to success.
As part of this study we would like to collect some information on the children in your child's class before the programme begins, and again at the end of the Reception year, to see how much their language and social skills improve. This will be compared with information from schools that did not receive the programme, to identify what difference it made. We would very much like to include your child in the study. More details are included in the 'further information over the page. **Researchers will not be evaluating your child as an individual and we will not use your child's name or the name of the school in the research reports**. We very much hope that you will allow your child to take part.
If you are happy for your child to take part, please complete and return the enclosed form. If we do not receive the form, your child's information will not be included in the study
With thanks from the Research and Evaluation Team

Education Endowment Foundation · University of Oxford · Institute of Education UCL · A+ Education Ltd · National Institute of Economic and Social Research · The Behavioural Insights Team

MORE DETAILED INFORMATION

Around 120 schools are taking part in the study, which is funded by the Education Endowment Foundation. Some schools will receive the training in 2017, and others will have the option to receive it from 2018 (this will be decided at random).

What information will you be collecting on my child?

Children's language and social skills will be tested at the start of the study - autumn 2016 - and again in summer 2018:

- language skills will be assessed using short tests conducted by trained assessors. The tests are designed to be fun and we expect your child will enjoy taking part. For example, s/he might be shown pictures of four things (a chair, T-shirt, spoon and apple) and asked to point to the apple. One of the tests will involve recording your child as they respond to prompts from a story;

- we will also ask your child's teacher to complete a questionnaire about his or her social development. For example, we will ask whether your child shares easily with other children and is confident with other people.

How will we use the information?

The information from these tests will be used by the research team, along with information provided by your child's school (e.g. your child's name, date of birth, gender, unique pupil number, whether English is your child's first language). They will use it to decide if the training for teachers has been successful in improving the quality of their practice and children's language and social development.

For the purposes of research, this information will also be linked with information about your child from the National Pupil Database (held by the Department for Education) and shared with the evaluation team, the research team, the Department for Education, EEF, EEF's data contractor FFT Education and (in an anonymised form) to the UK Data Archive.

Your child's data will be treated with the strictest confidence. Any paper or audio records will be stored in an anonymised form, and will be destroyed 24 months after the final report for this study is produced. You may withdraw your child from undertaking the tests, and/or withdraw any data relating to your child from the study, at any time.

Who do I contact for more information?

If you have any questions or would like any further information, please contact Daniel Carr: Daniel.Carr@behaviouralinsights.co.uk

This project has been approved by the University of Oxford Ethics Committee. If you would like more information about this, or have any concerns during the course of the research, please contact the Chair of Department of Education Research Ethics Committee, Dr Liam Gearon: liam.gearon@education.ox.ac.uk

Using Research tools to improve Language in the Early Years

Parent/Carer Consent Form

If you have read and understood the information provided, and are happy for your child to take part in the project, please complete and sign this form.

Child's name (BLOCK CAPITALS):
.....

Parent/carer name (BLOCK CAPITALS):
.....

Parent/carer signature:
.....
.....

Date: .....

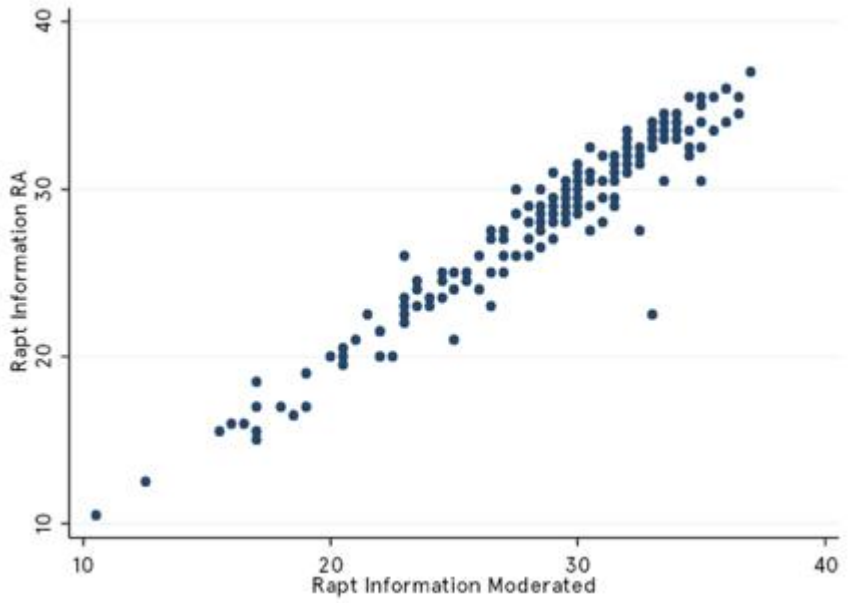
(Please return the completed form to your child's class teacher)

Appendix D: Moderation Analysis

In order to ensure that the scoring of outcome data was done to a reasonable standard of quality, we conducted exploratory moderation analysis. This exercise determines the extent to which a subset of scores, marked by the research assistants (RAs), differed when marked by a member of the Oxford University team.
For this analysis, the moderator marked 10% of the Renfrew Action Picture Test (RAPT) assessments, scoring both information and grammar outcomes. The moderator's scores were then compared with the scores given by the RAs. Our comparison consisted of 3 steps: (1) comparing means and standard deviations, (2) checking correlations between sets of scores, and (3) observing scatterplots.
Means and standard deviations proved to be extremely similar. For the RAPT information outcomes, the RA-marked mean and SD were 28.27 and 5.08, respectively, and in the moderated outcomes, the mean and SD were 28.73 and 4.97, respectively. For the RAPT grammar outcomes, the RA-marked mean and SD were 21.78 and 5.40, respectively, and in the moderated outcomes, the mean and SD were 22.93 and 5.36, respectively.

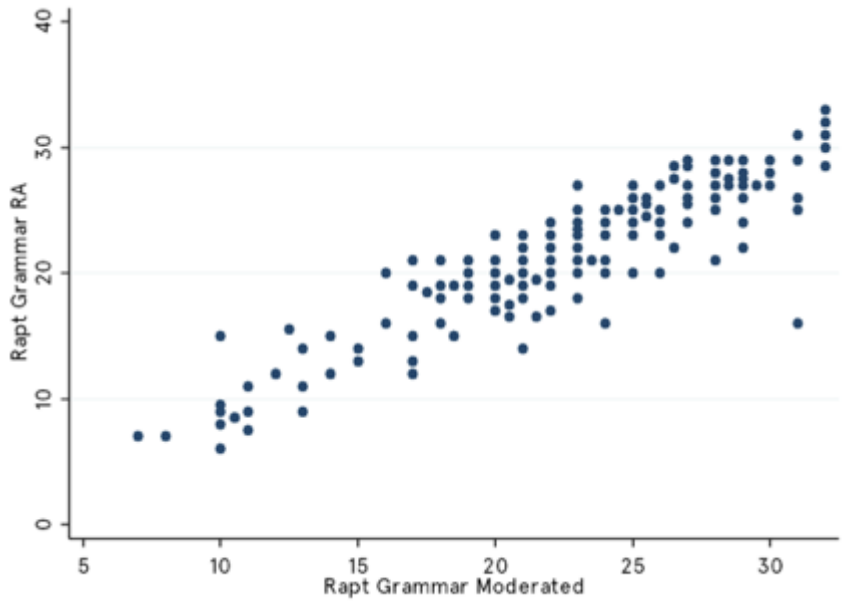
The correlation coefficient of RAPT information scores marked by RAs and moderated scores is 0.9635. Correlation coefficients above 0.9 suggest a high level of correlation between the two components. The scatterplot below depicts the strong correlation.

This scatterplot shows the correlation between RAPT information scores marked by the RAs and moderated RAPT information scores. The linear pattern of the plot suggests a strong correlation.



The correlation coefficient of RAPT grammar scores marked by RAs and moderated scores is 0.9046. The high degree of correlation is denoted by the correlation coefficient, which is above 0.9, and the linearity of the scatterplot below.

This scatterplot shows the correlation between RAPT grammar scores marked by the RAs and moderated RAPT information scores. The linear pattern of the plot suggests a strong correlation.



Based on the above evidence that the RA-marked scores closely resemble the moderated scores, we are confident that the scoring of outcome data was done to a sufficiently high standard.

Appendix E: Histograms of pre-test scores

The figures below present the distribution of pre-test scores and post-test scores in the treatment and control groups at the point of randomisation. Figures E.1 and E.2 presents the scores for the composite language measure (the primary outcome); Figures E.3-E.10 present scores for the component language measures.

Figure E.1: Pre-test composite language score, treatment and control groups

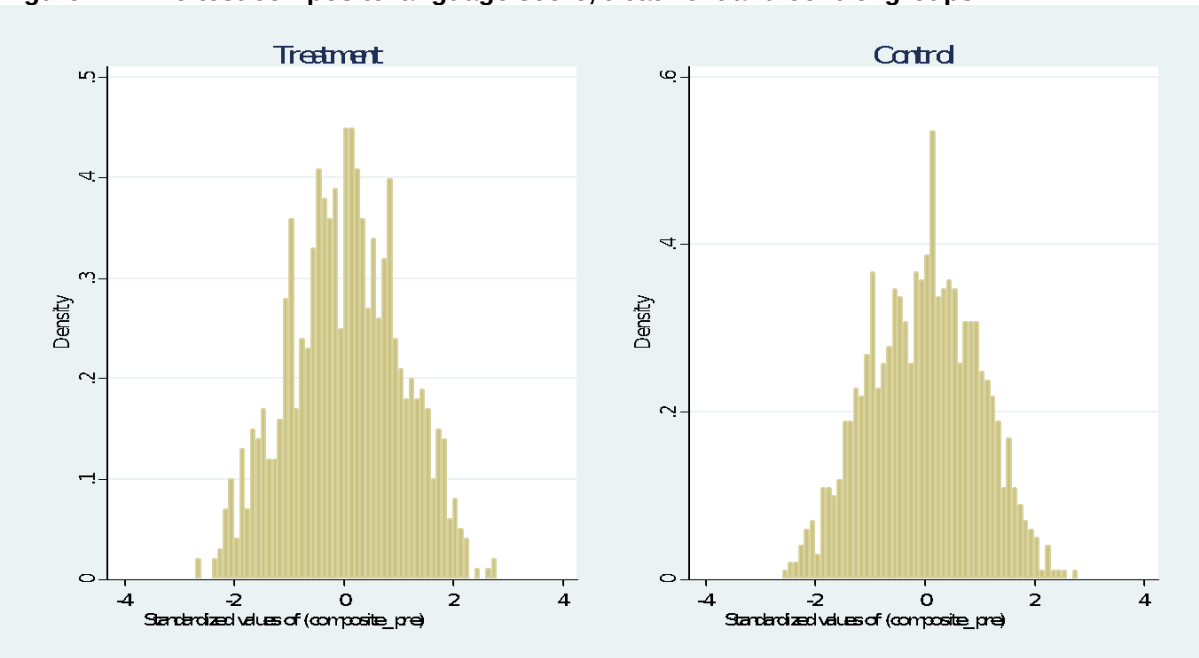


Figure E.2: Post-test composite language score, treatment and control groups

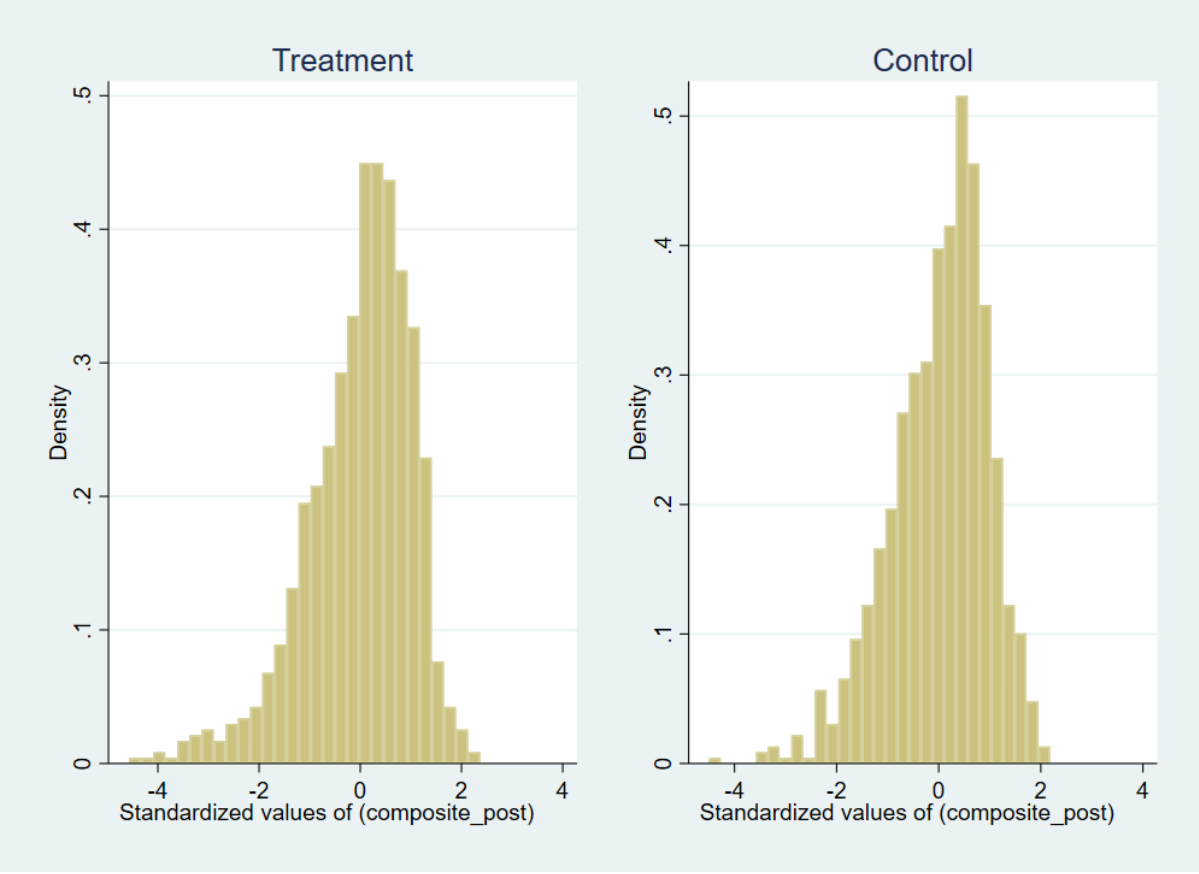


Figure E.3: Pre-test BPVS, treatment and control groups

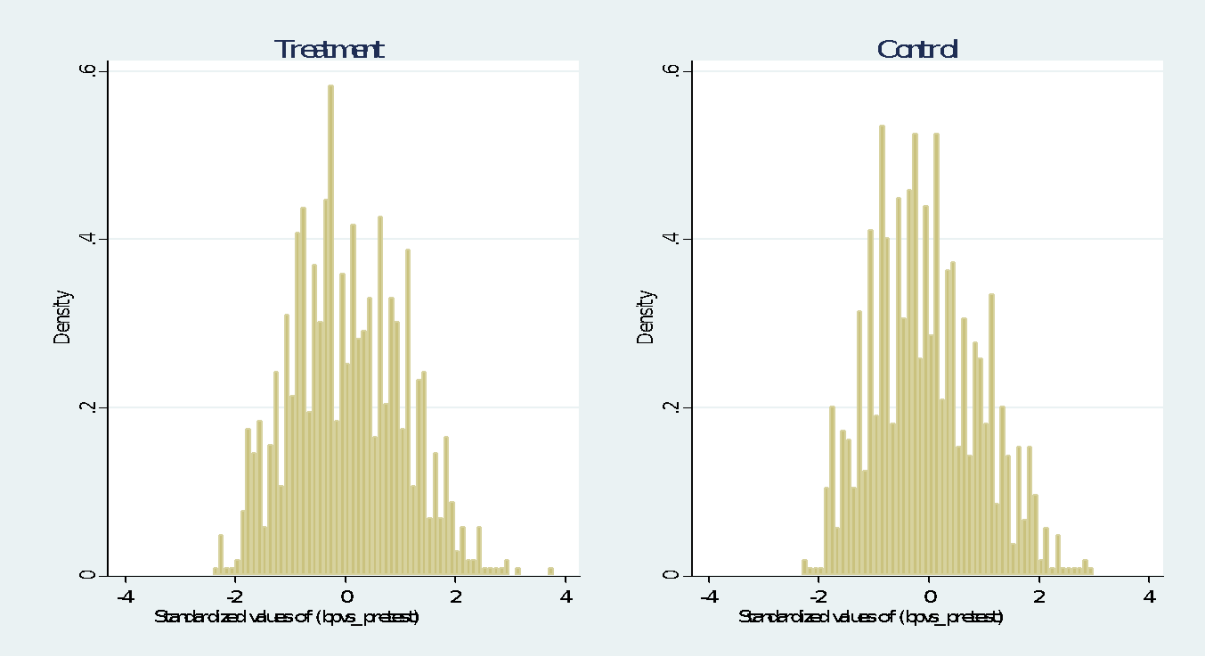


Figure E.4: Post-test BPVS, treatment and control groups

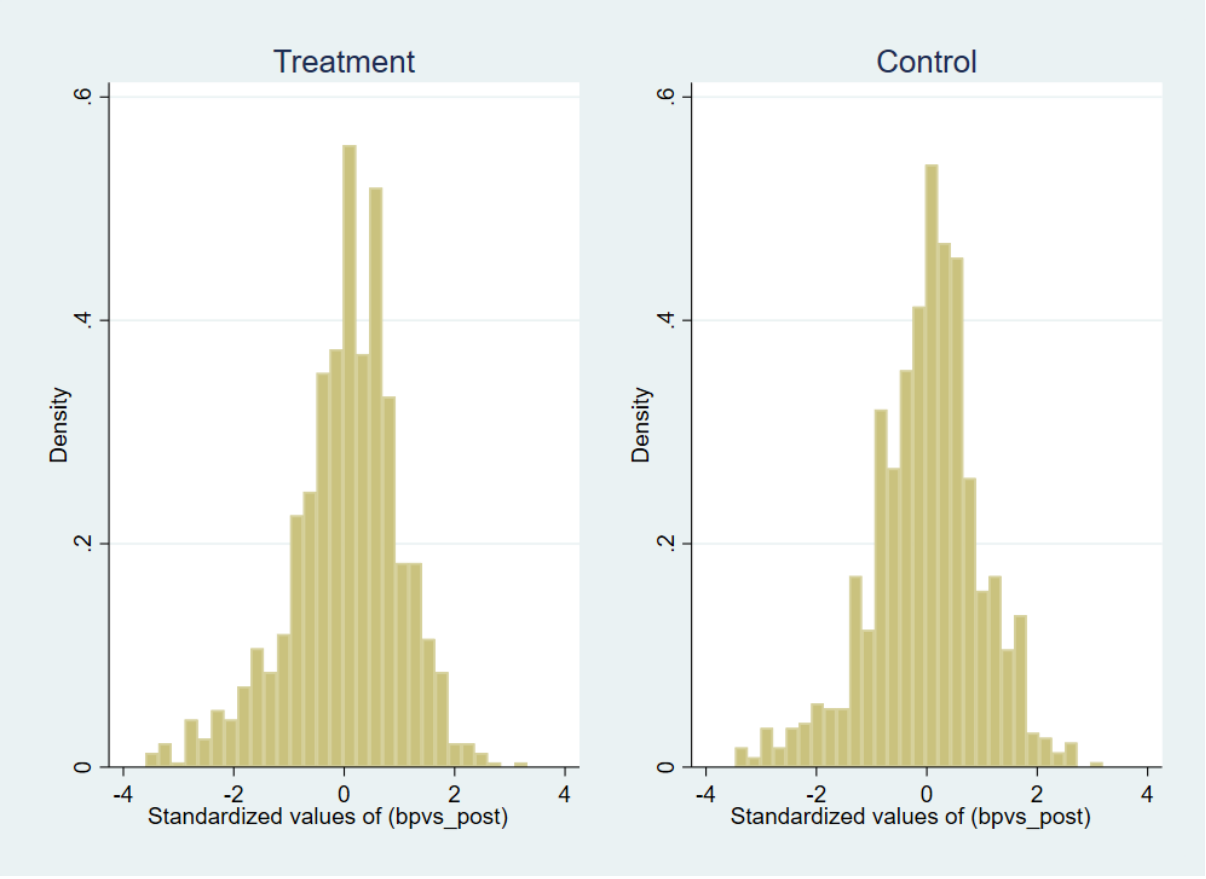


Figure E.5: Pre-test RAPT information score, treatment and control groups

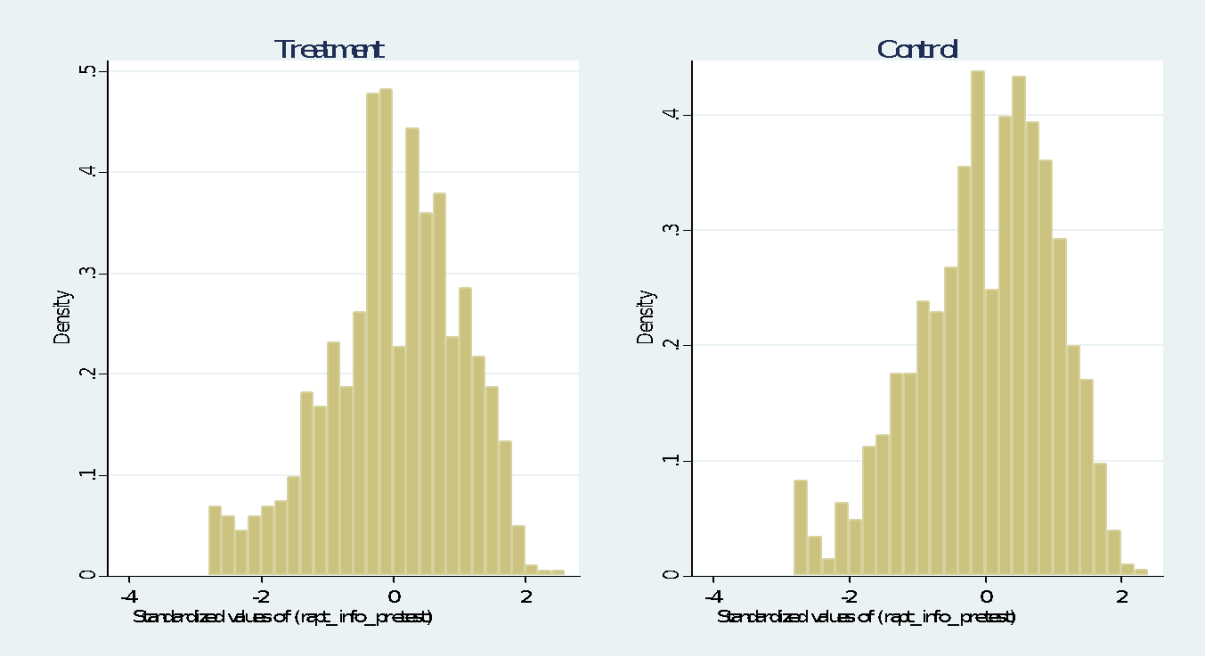


Figure E.6: Post-test RAPT information score, treatment and control groups

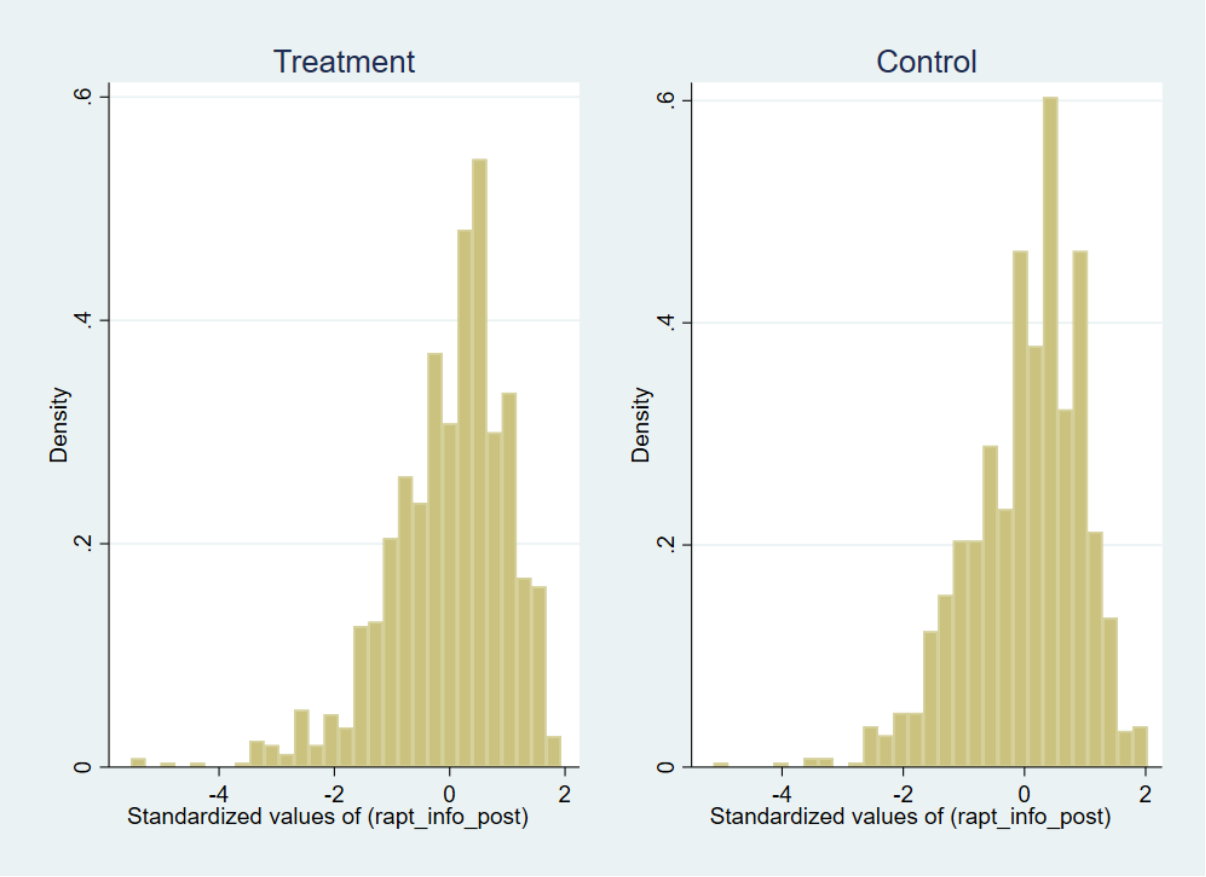


Figure E.7: Pre-test RAPT grammar score, treatment and control groups

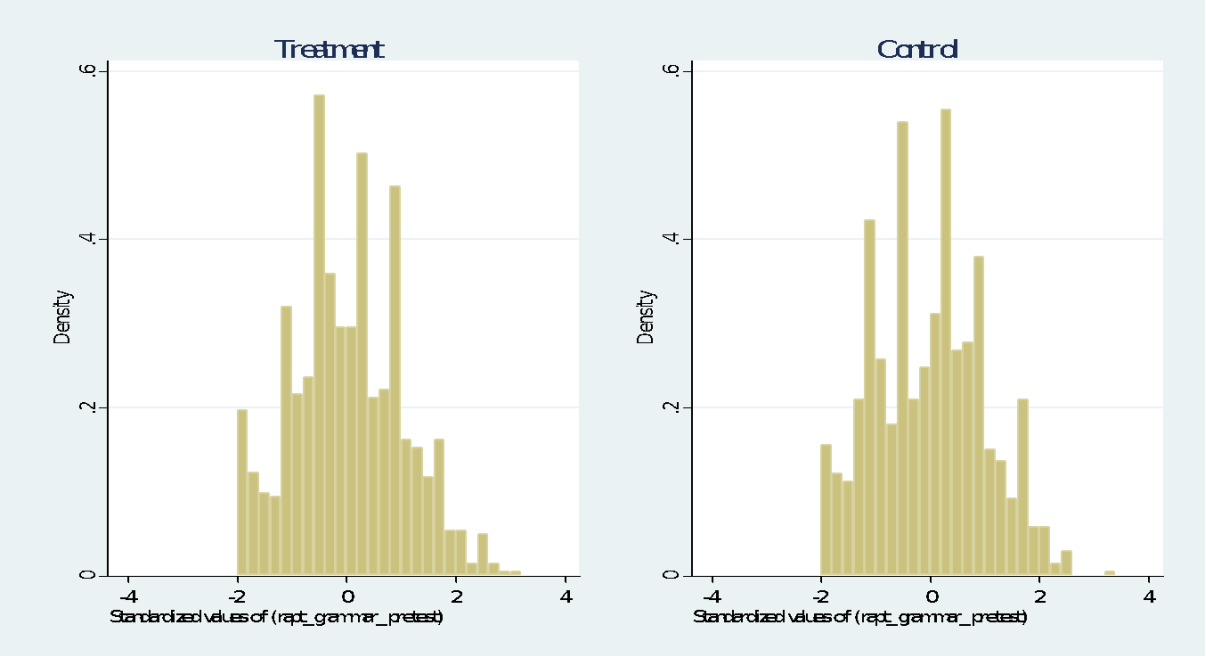


Figure E.8: Post-test RAPT grammar score, treatment and control groups

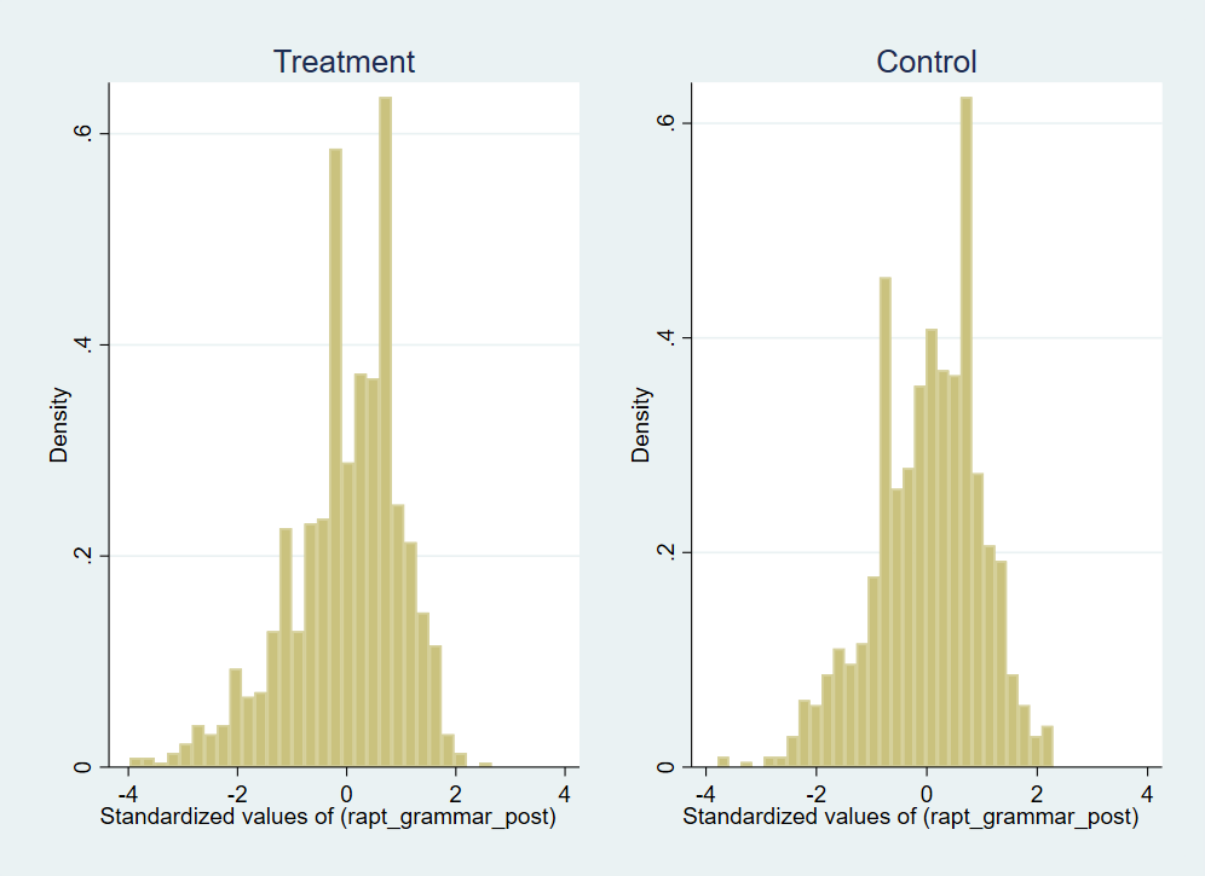


Figure E.9: Pre-test CELF, treatment and control groups

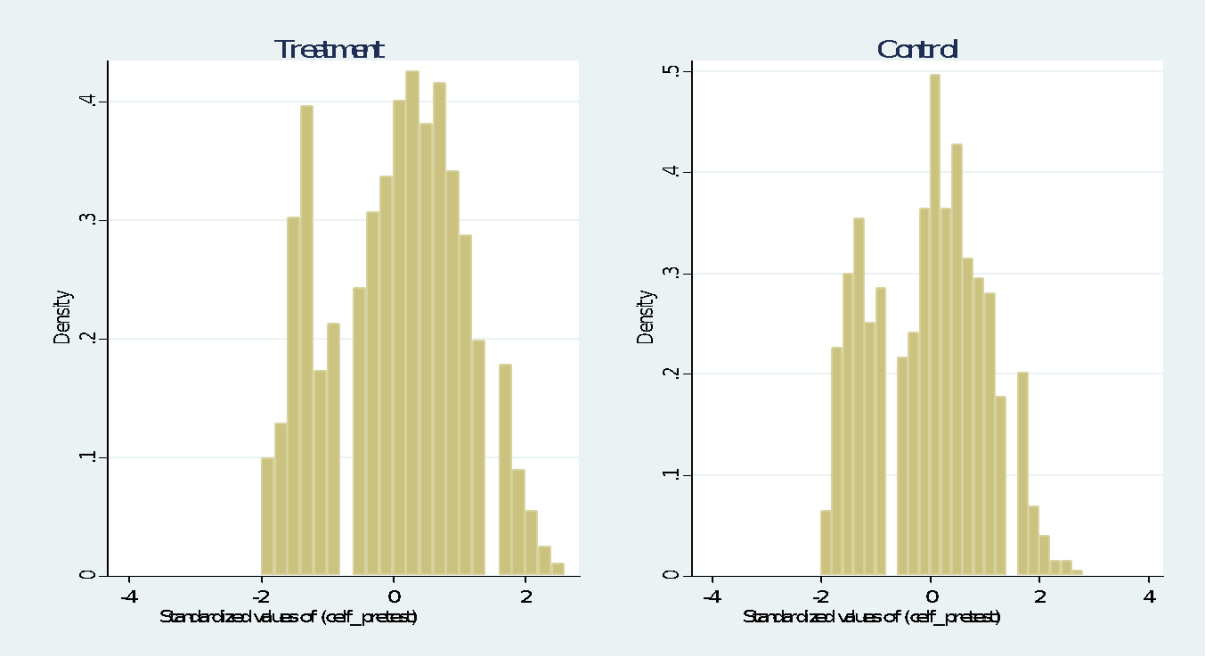
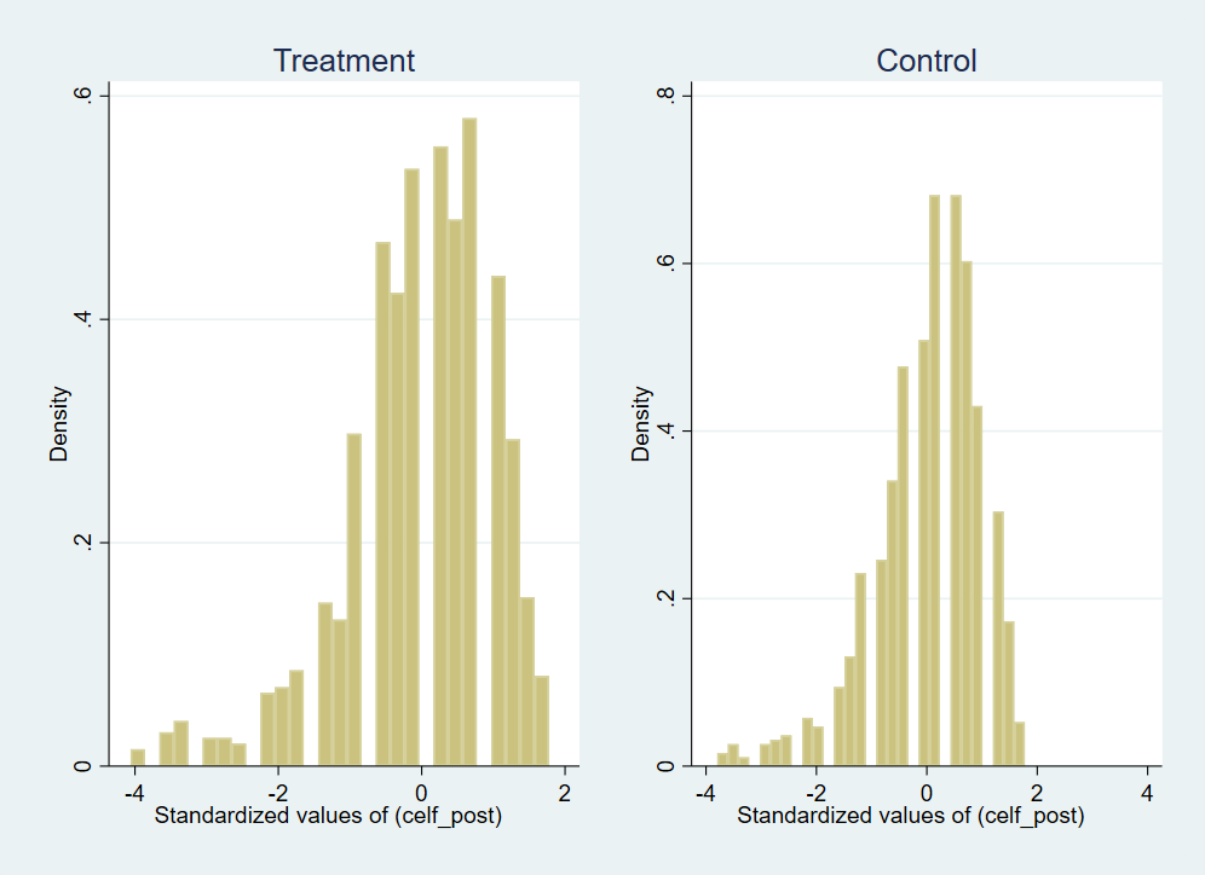


Figure E.10: Post-test CELF, treatment and control groups



Appendix F: Histograms – ASBI scores

Figure F.1: Total ASBI scores at follow-up, treatment and control groups

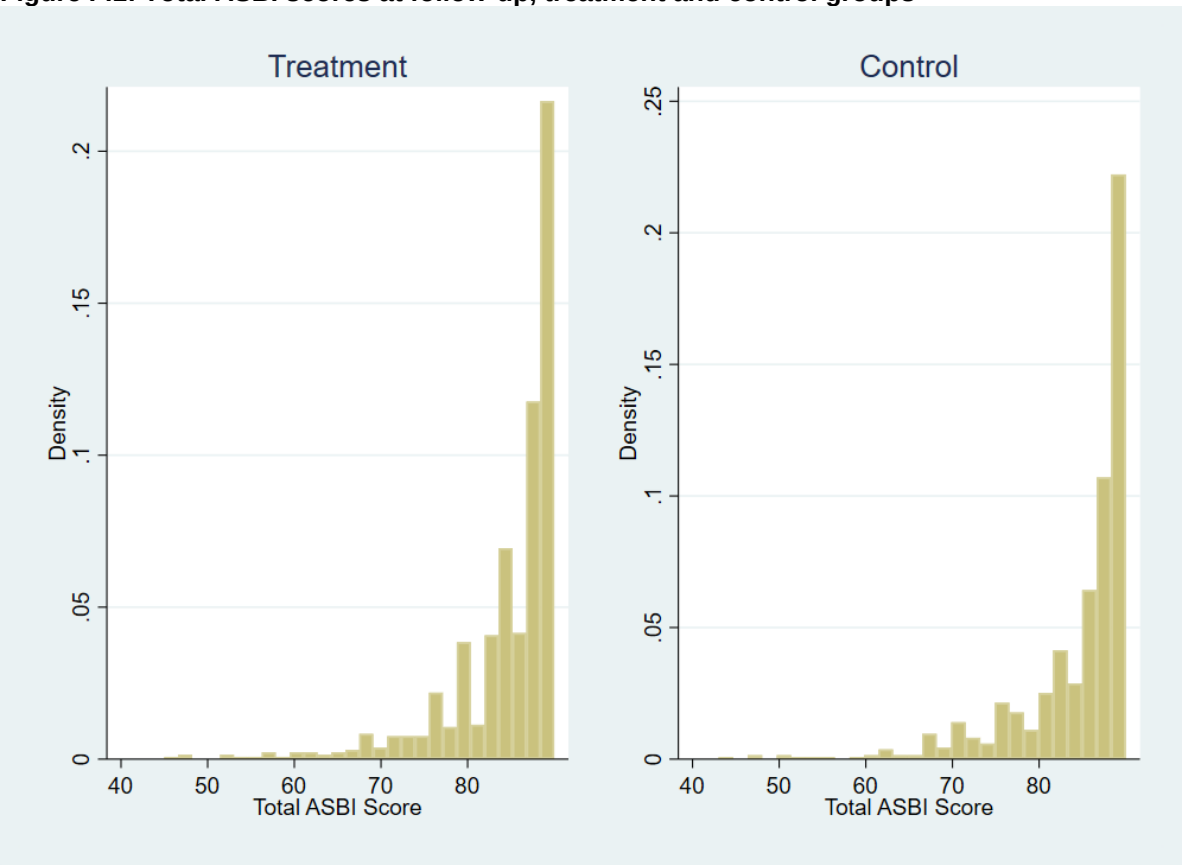


Figure F.2: Express ASBI scores at follow-up, treatment and control groups

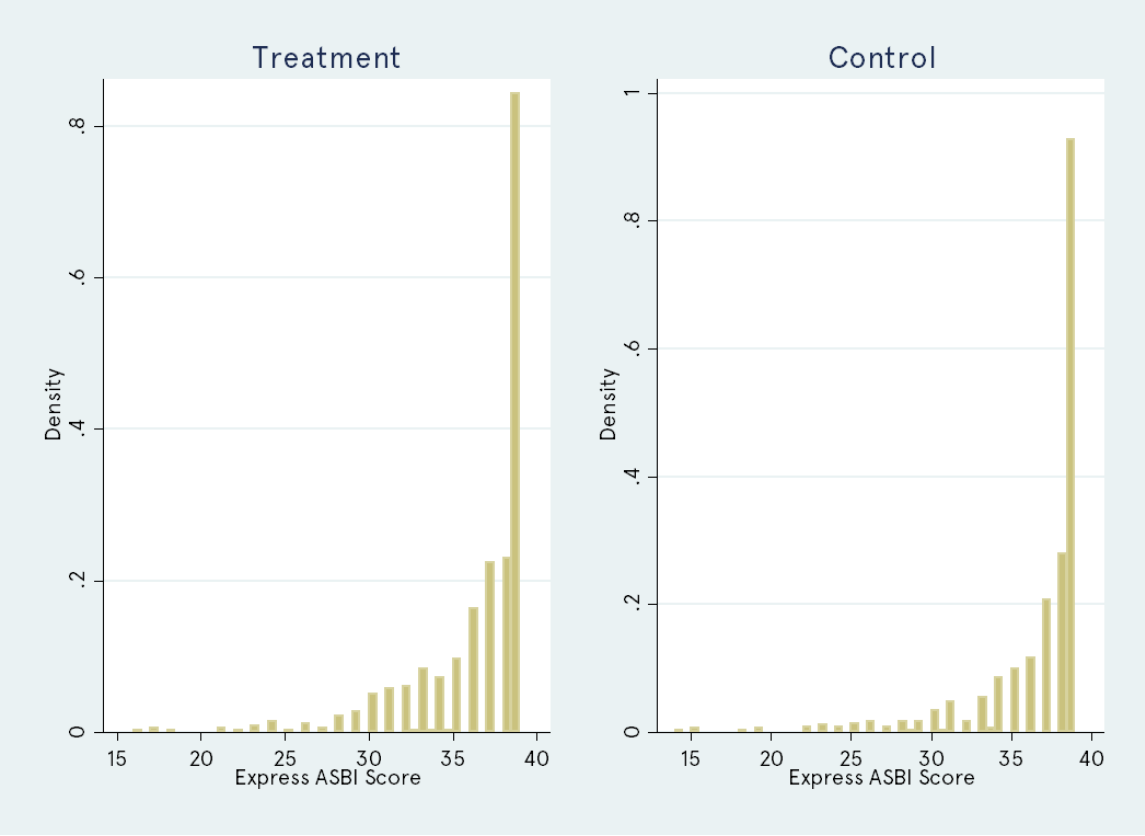


Figure F.3: Comply ASBI scores at follow-up, treatment and control groups

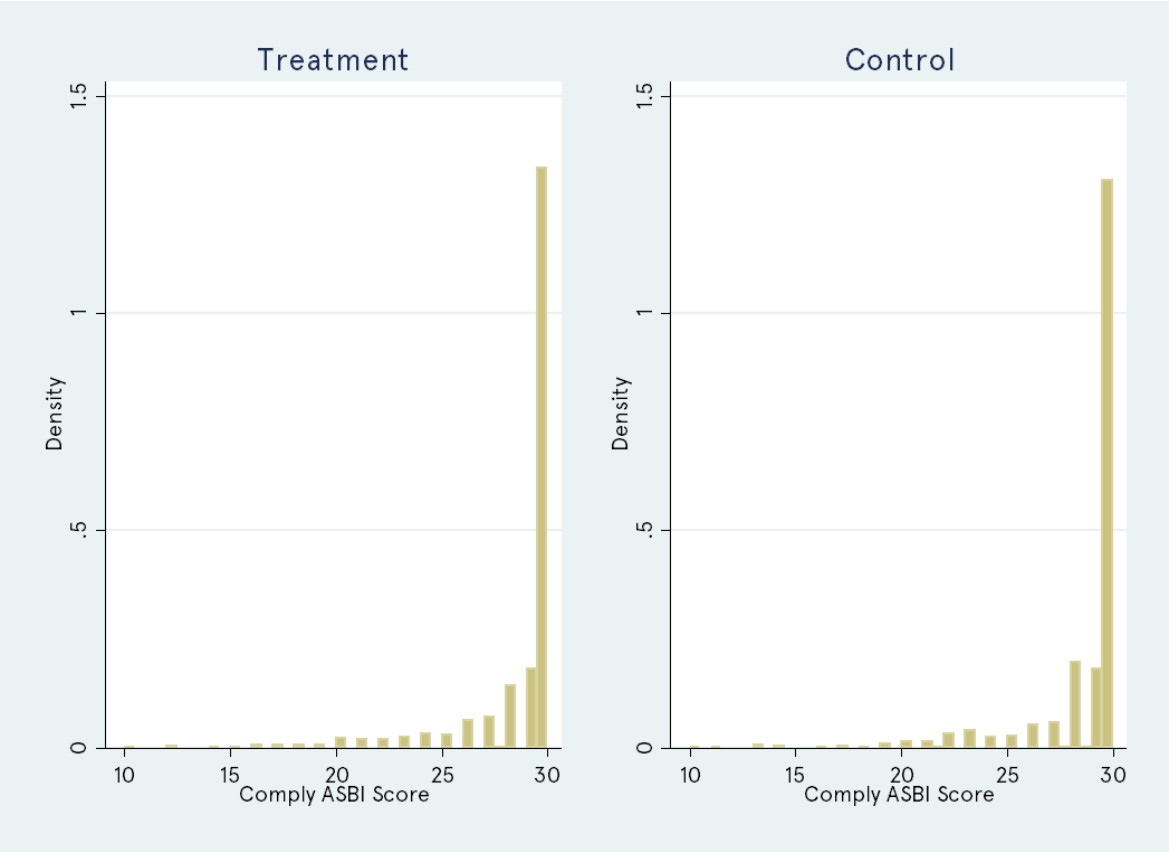


Figure F.4: Disrupt ASBI scores at follow-up, treatment and control groups

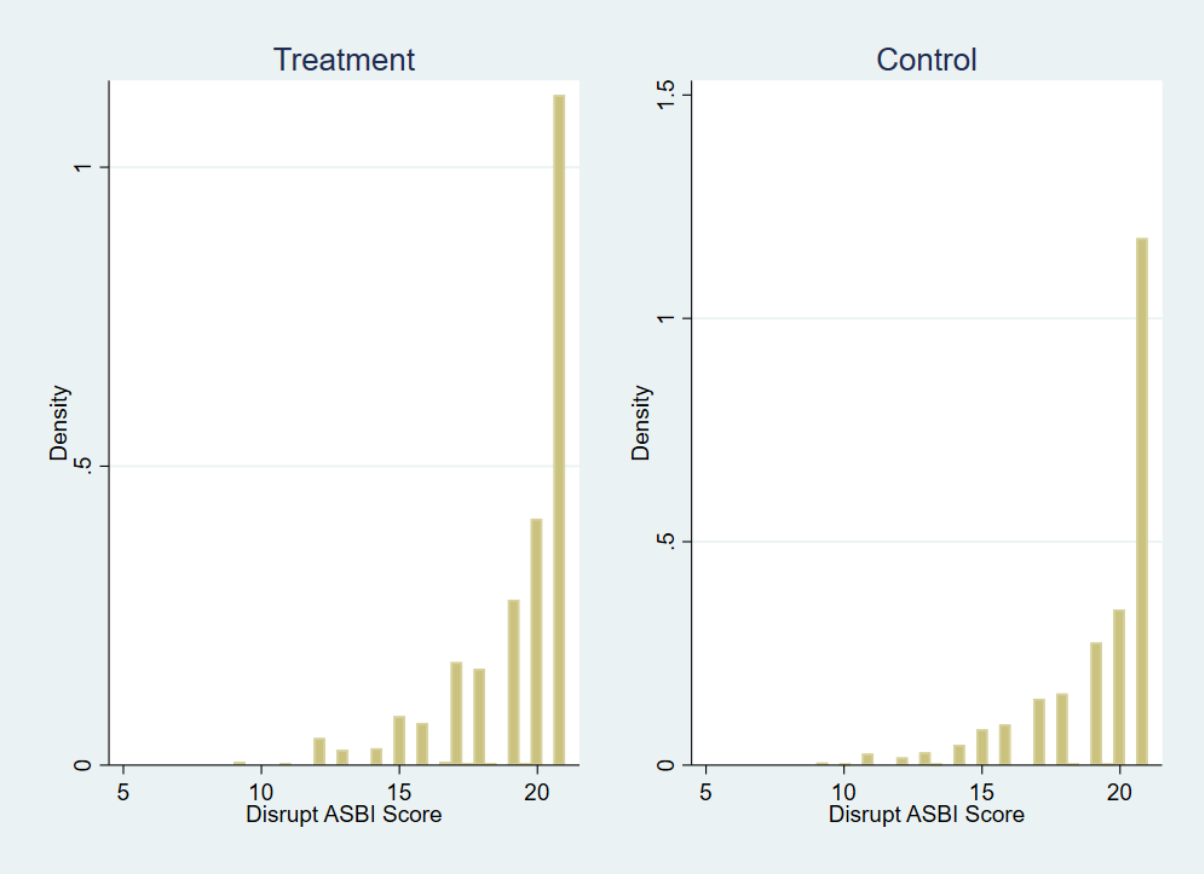
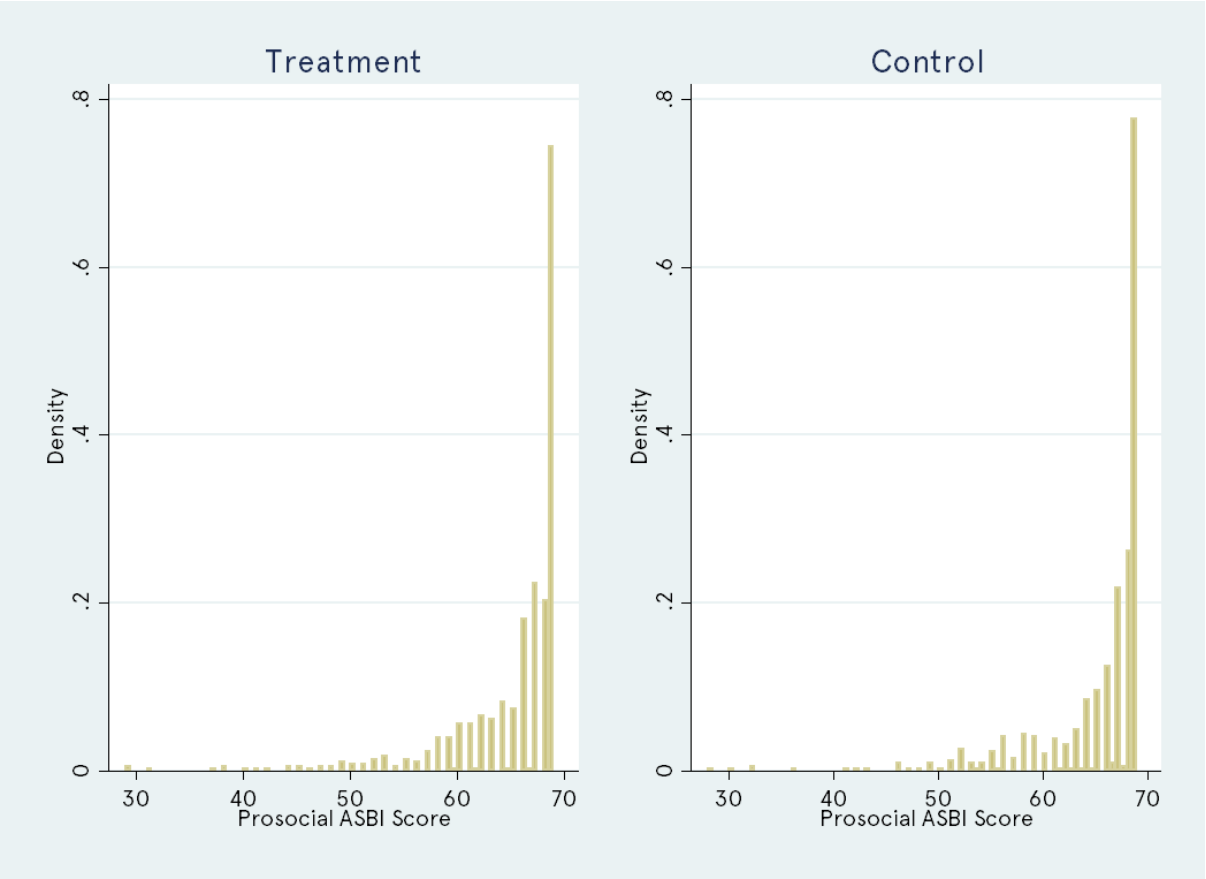


Figure F.5: Prosocial ASBI scores at follow-up, treatment and control groups


Treatment
Control
Density
0
.2
.4
.6
.8
30
40
50
60
70
Prosocial ASBI Score

Appendix G: Histograms – ERS scores

Figure G.1: Composite ERS scores at follow-up, treatment and control groups

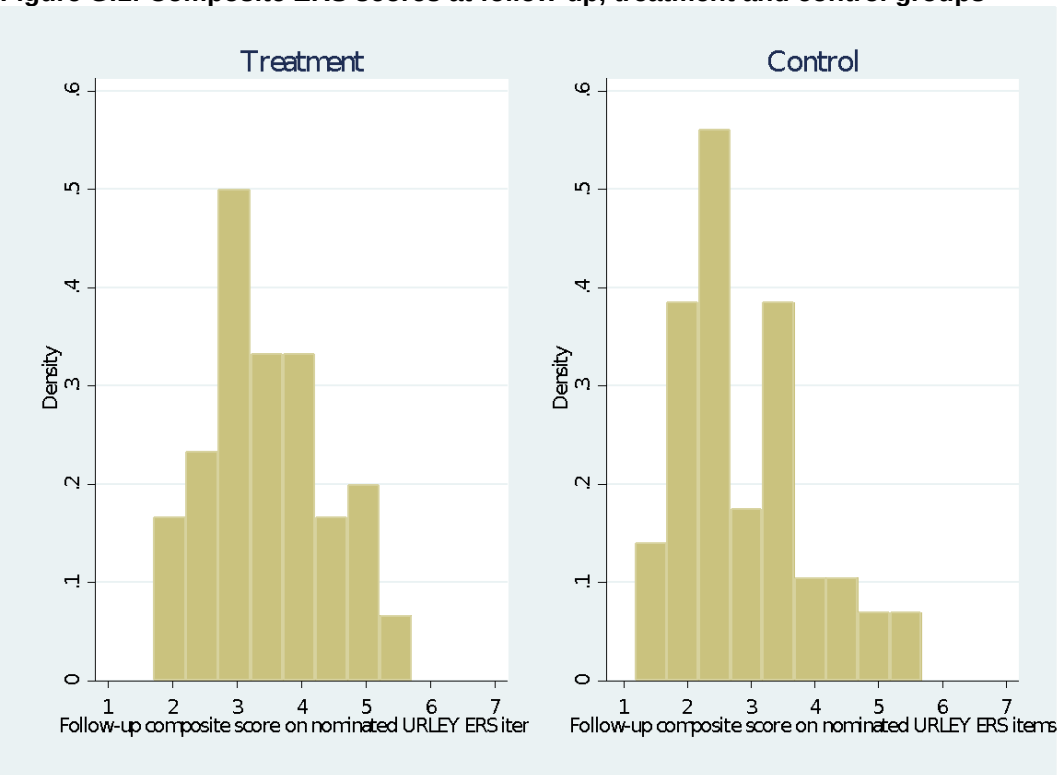


Figure G.2: ECERS-3 scores at follow-up, treatment and control groups

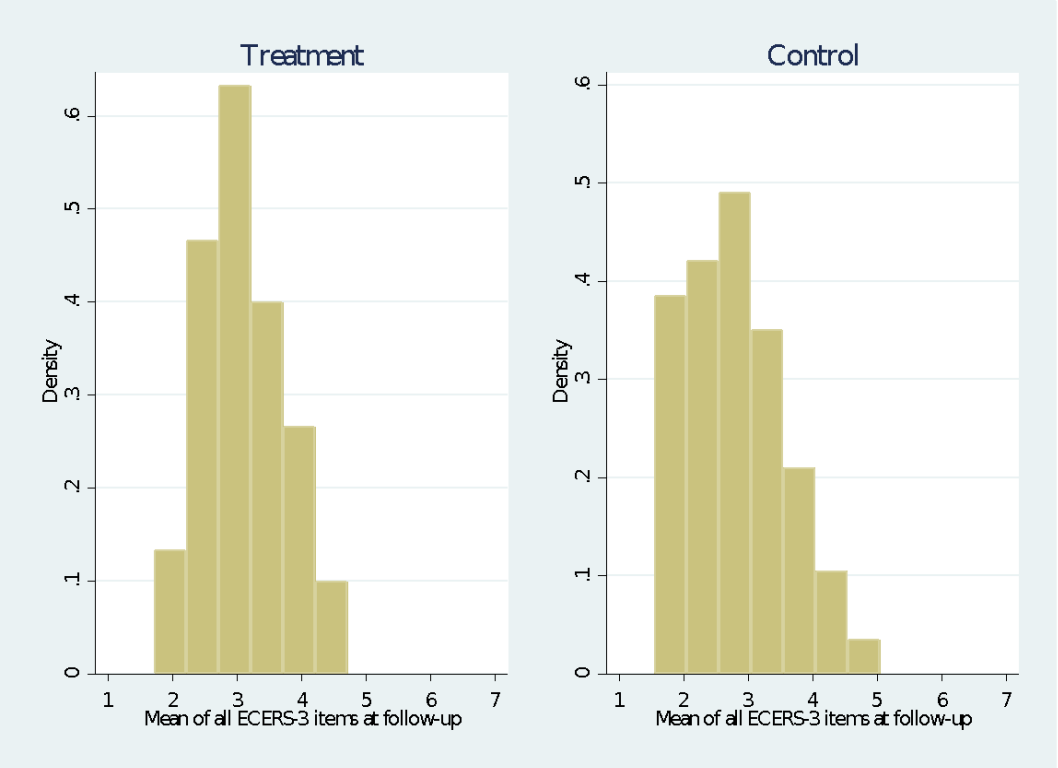


Figure G.3: ECERS-E scores (literacy subscale) at follow-up, treatment and control groups

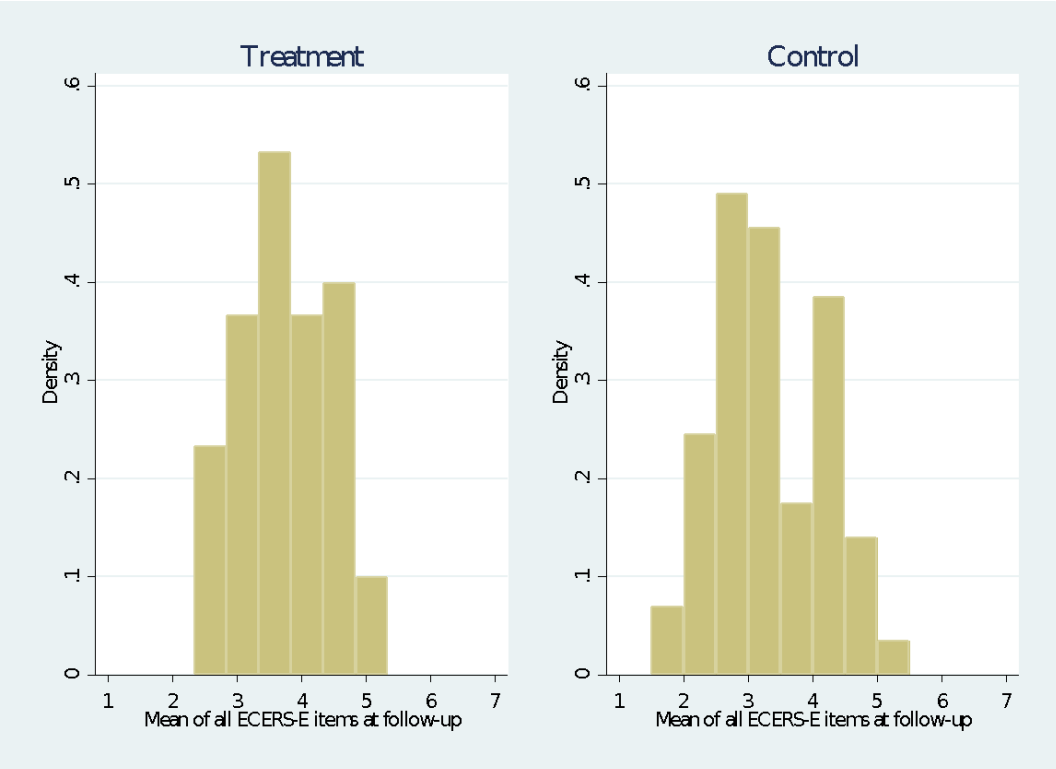
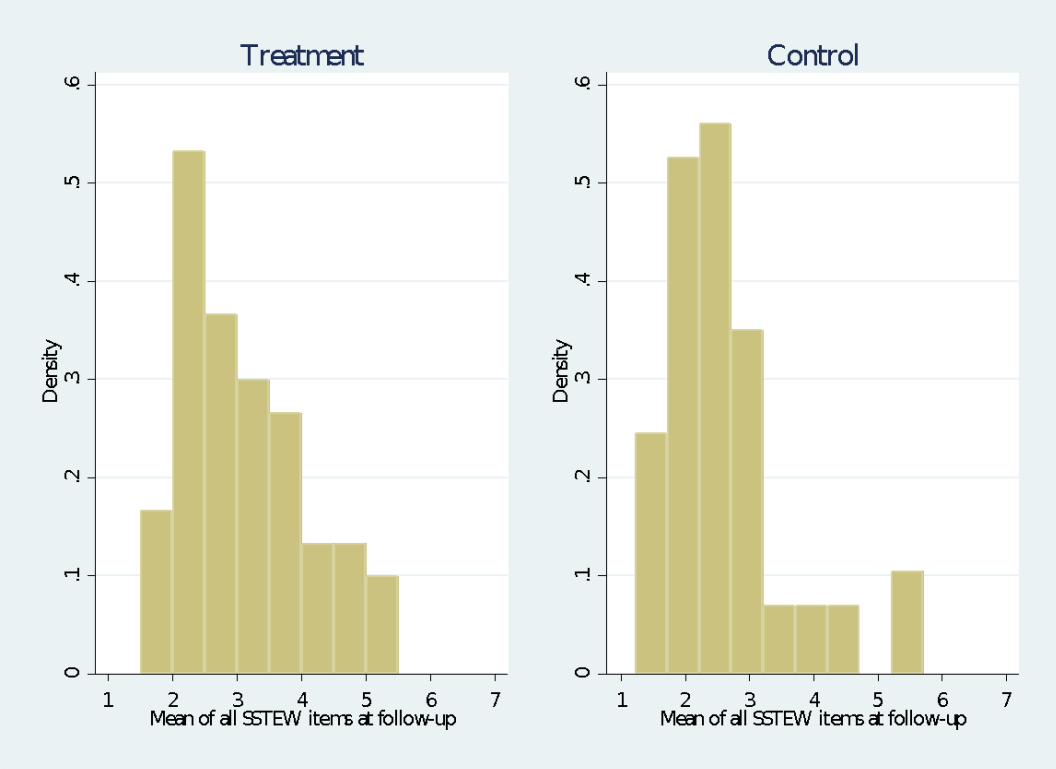


Figure G.4: SSTEW scores at follow-up, treatment and control groups



Appendix H: Additional analysis

Including pre-scores, restricting only to those with available pre- and post-scores the SAP prescribed additional analysis around missing data. We ran the primary analysis including only those with available pre- and post-test scores, replacing the school-level average pre-test with the individual pre-test scores. On this basis we still find a small negative effect size of the intervention on the composite language score; however, this difference is not statistically significant. When retaining more of the sample by including individual pre-test scores where available and then and imputing for those without pre-test scores, we again find a small, negative, but not statistically significant effect of the treatment on the composite language outcome. These results, which are consistent with the primary analysis, are reported in the first two columns of Table 26.

Table 26: Various specifications for composite language measure

	Including those with available pre- and post-scores	Including those with pre-scores and imputing those without pre-scores	Including FSM and EAL pupils as additional controls
Treatment	-0.053 [-0.151, 0.045]0.050)	-0.068 [-0.166, 0.03]	-0.024 [-0.139,0.092]
N	1604	1965	1976

Note: standard errors based on school-level clusters. 95% confidence intervals reported in brackets. Models also control for blocking dummies of strata used in randomisation and pre-test score. Statistical significance indicated as follows: * p<0.05; ** p<0.01; *** p<0.001

In the third column of Table 26, we report the results of running an additional model which additionally controls for whether pupils were eligible for FSM and for EAL pupils. This analysis was not pre-specified in the SAP, but we include it here due to the imbalance in these characteristics between treatment and control groups. In this specification, the effect size is slightly reduced at -0.02, equivalent to no additional (or less) progress.

Finally, we check the sensitivity of the results to the inclusion of pupils for whom some but not all language measures are available. Still, Table 27 shows that we find no statistically significant impact of the treatment on any of the four language measures.

Table 27: BPVS including those with pre- and post-scores

	BPVS	RAPT Info	RAPT Grammar	CELF
Treatment	-0.016 [-0.104, 0.072]	-0.030 [-0.153, 0.093]	-0.038 [-0.167, 0.091]	-0.063 [-0.179, 0.053]

N	1657	1633	1633	1615

Note: standard errors based on school-level clusters. 95% confidence intervals reported in brackets. Models also control for blocking dummies of strata used in randomisation and pre-test score. Statistical significance indicated as follows: * p<0.05; ** p<0.01; *** p<0.001

Missing data analysis

Table 28: Analysis of missing post-test data (composite language score), regression results

	Primary analysis specification	Including additional pupil-level characteristics
Treatment	**-0.023** **(0.036)**	**-0.023** **(0.036)**
School-level pre-test average	**0.015** **(0.040)**	**0.026** **(0.041)**
Individual pre-test score		
Female		**-0.016** **(0.017)**
Ever eligible for FSM		**0.033** **(0.020)**

EAL		**0.019** **(0.031)**
Pupil age (months)		**-0.002** **(0.002)**
N	**2,504**	**2,504**

Note: standard errors based on school-level clustered standard errors reported in parentheses. Models also control for blocking dummies of strata used in randomisation. Statistical significance indicated as follows: * p<0.05; ** p<0.01; *** p<0.001.

Appendix I: Mediation analysis

Table H.1: Whether change in ERS scores mediates the treatment effect, exploratory analysis

	Estimated change in outcome unexplained by change in ERS score in treatment schools	Estimated change in outcome explained by change in ERS score in treatment schools
Composite ERS	0.044 (0.055)	-0.042 (0.071)
ECERS-3	0.039 (0.077)	-0.022 (0.098)
ECERS-E (literacy subscale)	-0.017 (0.068)	-0.024 (0.083)
SSTEW	-0.006 (0.063)	0.006 (0.076)

Note: standard errors based on school-level clustered standard errors reported in parentheses. Models also control for blocking dummies of strata used in randomisation and pre-test score. Statistical significance indicated as follows: * p<0.05; ** p<0.01; *** p<0.001. Each model is based on 1,978 pupils.

Appendix J: Factor analysis

Table I.1: Factor loadings (pattern matrix) and unique variances for language measures

Variable	Factor	Uniqueness
BPVS post-score	0.704	0.505
RAPT Info post-score	0.790	0.376
RAPT Grammar post-score	0.791	0.374
CELF post-score	0.652	0.576

Table I.2: Factor predictions

Variable	Factor
BPVS post-score	0.243
RAPT Info post-score	0.326
RAPT Grammar post-score	0.329
CELF post-score	0.204

Table I.3: Primary regression model for factor analysis

	Z Factor Post
Treatment	-0.079 (0.059)
N	1978

Note: standard errors based on school-level clustered standard errors reported in parentheses. Models also control for blocking dummies of strata used in randomisation and pre-test score. Statistical significance indicated as follows: * p<0.05; ** p<0.01; *** p<0.001

Appendix K: Code

Code K.1: Randomisation

```
clear
set more off
cd "C:\Users\daniel.carr\Google Drive\EEF\URLEY Evaluation Project\Randomisation"

*import data
************************************************************

*VC schools - this is the list of schools in the trial
import excel "C:\Users\daniel.carr\Google Drive\EEF\URLEY Evaluation Project\Randomisation\All URLEY
schools.xlsx", sheet("allschools") cellrange(A1:E124) firstrow
rename _all, lower

rename laestabnumber laestab
rename postcode pcode

*note - sorting URN for all data files pre-merge as School Unique Reference Numbers will be used for matching in our
merge.
sort urn

save "URLEY randomise.dta", replace

clear

*KS1 data - this is a data file from DfE's School Performance Comparison website for 2014-15 KS2 students, that also
contains there KS1 results in Average Point Score (tks1aps) form.
import delimited "2014-2015-england_ks2final"
rename _all, lower

drop if estab==.

gen space = " "
egen laestab=concat(lea space estab)
label variable laestab "LA Establishment Number"
order laestab, after(estab)

keep urn schname laestab tks1aps urn_ac

replace urn_ac = . if urn_ac==0

destring tks1aps, replace force

destring urn, replace force
drop if urn==.

sort urn

save "DfE ks2.dta", replace

clear

*FSM data - this is also a data file from the above source, but here uses the school-wide census, from which we draw
the total proportion of 'FSM ever' students (those who've been on FSM, or are currently on FSM)
import delimited "2014-2015-england_census"
rename _all, lower
```

```
gen space = " "
egen laestab=concat(la space estab)
label variable laestab "LA Establishment Number"
order laestab, after(estab)

keep urn laestab pnumfsmever

destring pnumfsmever, replace force

drop if urn=="NAT"
destring urn, replace force
drop if urn==.

sort urn

save "DfE census.dta", replace

***** merge - original school list with the two DfE data sources to get our FSM and KS1 APS variables for the stratification.
clear
use "URLEY randomise.dta"
*match
merge 1:1 urn using "DfE census", gen(_censusmerge)
keep if _censusmerge==3

save "URLEY fsm merged.dta", replace

*all schools matched

*now merge on KS1 (10 won't match on URN, locate these first and try to match on LAESTAB)

merge 1:1 urn using "DfE ks2", gen(_KS2merge)
keep if _KS2merge==1
keep schoolname laestab pcode randomisationround pnumfsmever _censusmerge
merge 1:1 laestab using "DfE ks2", gen(_altmerge)
drop if _altmerge==2
save "matched on LAESTAB.dta", replace

*2 of 10 were matched on this basis - so 8 not matched overall

clear

use "URLEY fsm merged.dta"
merge 1:1 urn using "DfE ks2", gen(_KS2merge)
keep if _KS2merge==3
append using "matched on LAESTAB"

drop _censusmerge schname _KS2merge _altmerge

save "URLEY merged.dta", replace

**** create stratification variables

*FSM
egen fsmmed = median(pnumfsmever)

generate fsm = 1 if pnumfsmever > fsmmed
replace fsm = 0 if pnumfsmever <= fsmmed
replace fsm = . if pnumfsmever ==.

label def fsm 1 "High share" 0 "Low share"
label val fsm fsm
```

```
*KS1
egen ks1med = median(tks1aps)

generate ks1 = 1 if tks1aps > ks1med
replace ks1 = 0 if tks1aps <= ks1med
replace ks1 = . if tks1aps ==.

label def ks1 1 "High" 0 "Low"
label val ks1 ks1

*areas
sort pcode
gen area = 3
label def area 1 "West Midlands" 2 "Liverpool" 3 "Manchester"
label val area area

gen firsttwo = substr(pcode,1,2),a(pcode)
replace area = 2 if firsttwo =="WA"|firsttwo =="CH"|firsttwo =="L1"|firsttwo =="L2"|firsttwo =="L3"|firsttwo
=="L4"|firsttwo =="L5"|firsttwo =="L6"|firsttwo =="L7"|firsttwo =="L8"|firsttwo =="L9"
replace area = 1 if firsttwo =="WS"|firsttwo =="WV"|firsttwo =="B1"|firsttwo =="ST"|firsttwo =="DY"|firsttwo
=="B2"|firsttwo =="B3"|firsttwo =="B4"|firsttwo =="B5"|firsttwo =="B6"|firsttwo =="B7"|firsttwo =="B8"|firsttwo =="B9"

drop fsmmed ks1med firsttwo

save "URLEY for randomise.dta", replace

**** randomise batch 1

drop if randomisationround==2
*include the three stratifying variables, and use the missing command. Ratio is 1:1.
bitrandomise area fsm, gen(allocation) seed(14431) groups(1 1) missing

label def allocation 1 "Control" 2 "Treatment"
label val allocation allocation
label variable allocation "Trial Arm Allocation"

sort allocation

drop randomisationround pnumfsmever tks1aps fsm ks1 urn_ac

export excel using "EEF URLEY Batch 1 FINAL", sheetreplace firstrow(varlabels)


*** randomise batch 2
use "URLEY for randomise"
drop if randomisationround==1

*one school dropped out
drop if schoolname=="Rimrose Hope CE Primary School"

*include the three stratifying variables, and use the missing command. Ratio is 1:1.
bitrandomise area fsm, gen(allocation) seed(1443441) groups(1 1) missing

label def allocation 1 "Control" 2 "Treatment"
label val allocation allocation
label variable allocation "Trial Arm Allocation"

sort allocation

drop randomisationround pnumfsmever tks1aps fsm ks1 urn_ac
```

export excel using "EEF URLEY Batch 2 Provisional", sheetreplace firstrow(varlabels)

Code J.2: Analysis

Our primary analysis is run using the following code:

regress z_composite_post allocation dblock2-dblock12 avg_pre_comp_score, vce(cluster schoolcode)

where:

z_composite_post is our primary outcome (composite language score)

allocation indicates whether the pupil is in the treatment or control group

avg_pre_comp_score is the school-level average pre-test composite language score (our measure of prior attainment)

dblock* are the stratum indicators used in the randomisation

Our secondary analysis is run using the following code:

```
local outcomes z_bpvs z_rapt_info z_rapt_grammar z_celf
	foreach measure of local outcomes {
		regress `measure'_post allocation dblock2-dblock12 avg_`measure'_pretest_score, /// vce(cluster schoolcode)
	}
*ASBI analysis - same specification as above
	foreach asbi in totalasbi express comply disrupt prosocial {
		regress `asbi'_post allocation dblock2-dblock12 avg_`asbi'_pre_score, vce(cluster /// schoolcode)
```

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

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

https://educationendowmentfoundation.org.uk

@EducEndowFoundn

Facebook.com/EducEndowFoundn