ASHA American Speech-Language-Hearing Association

AJSLP

Research Article

Real-World Effectiveness of Early Language Intervention: Evidence From a Nationwide Rollout of the Nuffield Early Language Intervention in England

Gillian West,[a,b] Rocío Titiunik,[c] Sarah Hearne,[b,d] Caroline Korell,[d] Mihaela Duta,[d] and Charles Hulme[b,d,e]

[a] University College London, United Kingdom [b] OxEd and Assessment Ltd, Oxford, United Kingdom [c] Princeton University, NJ [d] University of Oxford, United Kingdom [e] Oxford Brookes University, United Kingdom

ARTICLE INFO

Article History:
Received May 18, 2025
Revision received August 26, 2025
Accepted November 26, 2025

Editor-in-Chief: Amy L. Donaldson
Editor: Michelle Therrien

https://doi.org/10.1044/2025_AJSLP-25-00205

ABSTRACT

Purpose: Oral language provides the foundation for learning and education, yet many children enter school with poor language skills, placing them at risk of long-term educational disadvantage. The Nuffield Early Language Intervention (NELI) is a targeted 20-week language program, designed to improve the language skills of children entering school with oral language weaknesses. While NELI has repeatedly demonstrated effectiveness in randomized controlled trials (RCTs), this study investigates whether those benefits are retained when the program is implemented at scale in diverse, real-world educational settings.
Method: A rollout of NELI was funded as part of the U.K. government's COVID-19 response and reached over 10,000 schools in England. To support effective implementation at this scale, key program components were adapted in line with principles from Damschroder's Consolidated Framework for Implementation Research. Adaptations included the development of asynchronous online training for school staff, sustained practitioner support, and the use of an automated language assessment to identify children to receive intervention and monitor their language development.
Results: A regression discontinuity (RD) design was used to estimate the causal impact of NELI by comparing children just above and below the threshold on the language assessment used to allocate the intervention. This quasi-experimental approach enabled causal inference without random assignment. Analysis of 19,936 children showed that those assigned to NELI made significantly greater progress in language development than those who were not (Hedges's $g = .40$).
Conclusions: This study is methodologically unusual in providing converging evidence from an RD analysis for an intervention previously demonstrated to be effective in a series of RCTs. The findings demonstrate that NELI can be successfully implemented at scale and lead to meaningful improvements in children's oral language skills. For educators and speech-language pathologists, this evidence supports the wider adoption of NELI as a practical, evidence-based intervention to address children's early language weaknesses in school.

Correspondence to Gillian West: g.west@ucl.ac.uk. ***Disclosure:*** *Charles Hulme and Gillian West are Directors of OxEd and Assessment Ltd, a University of Oxford spin-out company founded to distribute LanguageScreen as a commercial product. Mihaela Duta is a shareholder in OxEd and Assessment Ltd. Charles Hulme is an author of the Nuffield Early Language Intervention program, which is licensed to and distributed by OxEd and Assessment Ltd. All other authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

Oral language skills form the basis for educational success (Hjetland et al., 2020; Hulme et al., 2015), yet many children begin school with poor language skills. Among these are the approximately 7% of children who have developmental language disorder (Norbury et al., 2016), a group likely to require long-term, specialized support. Children from socioeconomically disadvantaged backgrounds also tend to enter school with language skills that are less well developed than more advantaged peers (Gilkerson et al., 2017; Hart & Risley,

1995; Norbury et al., 2021; Pace et al., 2017), and those with English as an additional language (EAL) often encounter compounded demands due to limited exposure to English before school (Hessel & Strand, 2023; Lester et al., 2025; Strand & Hessel, 2018). These diverse subgroups reflect a range of language needs, highlighting the importance of early support to improve language skills and identify children who may require further assistance. This can be achieved within a multitiered system of support (MTSS) framework that can match intervention intensity to level of need.

Targeted language interventions delivered by educators in schools (Tier 2 interventions) have been shown to be effective in improving children's language skills (Hulme et al., 2020; Snowling et al., 2022). However, it remains unclear how well interventions demonstrating effectiveness under trial conditions perform once they transfer into practice and what factors determine success or failure. This gap between research and practice has been termed the "last mile" problem (Gaias et al., 2023), and a growing transdisciplinary field of implementation science in education seeks to understand the critical factors needed for interventions to work at scale. A variety of conceptual frameworks have been developed to guide scaling work. These include the Consolidated Framework for Implementation Research (CFIR; Damschroder et al., 2009, 2022), which was originally developed within the context of clinical and health care settings. More recent frameworks, such as the Generic Implementation Framework for school-based research (Komesidou & Hogan, 2023), have since been developed specifically for educational contexts.

One intervention developed to address early language needs is the Nuffield Early Language Intervention (NELI). NELI is a fully manualized 20-week oral language intervention for children with weak language skills in their first year in school (Fricke et al., 2018), which in the U.K. context corresponds to children aged 4–5 years. It is a pull-out intervention, comprising both small-group and individual sessions, delivered by a trained teaching assistant. Sessions are scripted, targeting key domains of oral language development: vocabulary, narrative, and listening skills. Vocabulary is taught using a multicontextual approach within a repetitive framework. Narrative activities familiarize children with story structure and promote the use of newly learned vocabulary in connected speech, while listening work targets receptive language skills and phonological awareness.

Two efficacy trials and a large-scale effectiveness trial (Fricke et al., 2013, 2017; West et al., 2021) have shown that NELI is effective in producing educationally significant improvements in children's language skills. The first efficacy trial (Fricke et al., 2013) demonstrated substantial gains in vocabulary, narrative skills, and listening comprehension for children identified as having weak oral language skills at school entry, with effects sustained for several months after the intervention. A subsequent replication and extension study (Fricke et al., 2017) confirmed these findings in a larger and more diverse sample, showing that the program could be delivered effectively by trained teaching assistants. Most recently, a national-scale effectiveness trial (West et al., 2021) evaluated NELI under real-world implementation conditions in 193 schools. This trial provided robust evidence that NELI's benefits generalize to routine school practice. This trial also represented the first use of an automated, tablet-based screening and assessment app for children's oral language skills, LanguageScreen, to identify eligible children and measure outcomes. Taken together, these studies provide a strong and consistent body of evidence that NELI reliably produces meaningful improvements in oral language development.

In 2020, following the success of the effectiveness trial (West et al., 2021) funded by the Education Endowment Foundation (EEF), the United Kingdom's equivalent of the What Works Clearinghouse, the U.K. Department for Education (DfE) funded a national rollout of NELI as part of the government's COVID-19 recovery strategy. Aiming to address the impact of school closures on young children's language development, the program was offered to over 10,000 primary schools across England.

To ensure successful delivery across diverse school contexts, the rollout was informed by implementation science principles, particularly the CFIR (Damschroder et al., 2009). The CFIR offers a structured approach to understanding the factors that influence implementation success, recognizing that successful implementation depends on aligning the intervention with the multiple, nested layers of the educational environment, from individual teaching assistants and classroom practices to school leadership and national policy. It identifies five core domains that influence implementation success: intervention characteristics, inner setting, outer setting, characteristics of individuals, and process. These domains helped shape both the design and delivery of the NELI rollout. In addition to the structure and content of the intervention itself, three other factors were identified as important for implementation success: (a) the need for asynchronous online training, (b) ongoing just-in-time support for educators, and (c) provision of an automated language assessment (Hulme et al., 2024) to identify those children who most needed the NELI program and to measure the resulting benefits of the program.

The central research question in this study is: Does an evidence-based intervention such as NELI retain its effectiveness when scaled up and implemented in diverse, real-world school contexts? Answering this question is not straightforward, however. In any nonexperimental rollout of a policy intervention such as NELI, by definition, we

do not have intervention and control groups formed by random assignment. In a randomized controlled trial (RCT), random assignment of participants to trial arms equalizes both measured and unmeasured participant and school characteristics, effectively removing confounding variables (Shadish, 2002). Without the benefit of a randomly allocated control group to represent the counterfactual condition of "No intervention" in the rollout, analysis can only be quasi-experimental. However, carefully selected and well-specified quasi-experimental methods can provide a good substitute for randomization. A regression discontinuity (RD) analysis (Bloom, 2012) is one such quasi-experimental method that can be used to compare the performance of children receiving an intervention with a comparison of peers not receiving intervention. An RD design assigns participants to intervention and control conditions based on a known cutoff score on a pre-intervention score that usually measures need or merit (Cattaneo et al., 2020; Schochet et al., 2010). The assignment of the intervention follows a thresholding rule: All individuals whose scores are equal to or above the cutoff are assigned to the intervention condition, while all individuals whose scores are below the cutoff are assigned to the control condition. The RD design meets the prerequisites for establishing causal relationships (Institute of Education Sciences (Gleason et al., 2012).

The use of an automated language screening app, LanguageScreen (Hulme et al., 2024), provided a valid and reliable measure of the need for language intervention and a measure of language improvement for use in the RD analyses presented here. In our RD design, children's LanguageScreen scores at baseline (which we refer to as T1 or pretest LanguageScreen) are used for assignment to NELI, and LanguageScreen scores after the intervention (which we refer to as T2 or posttest LanguageScreen) is the outcome. Our analyses first replicate findings from an independent evaluation conducted on a subset of LanguageScreen data from the rollout (Smith et al., 2023) and then extend the analysis to include many more schools, giving a substantially larger sample size.

Understanding whether interventions remain effective when delivered at scale in real-world settings is critical to bridging the gap between research and practice. There are few studies that conduct statistically robust quantitative evaluations at the point of large-scale rollout. This study contributes to a small but growing body of implementation research (e.g., Pion & Lipsey, 2021), examining fully developed programs delivered in real-world school settings, using rigorous quasi-experimental methods to assess effectiveness beyond controlled trial conditions.

Method

As a result of COVID-19–related disruption to children's language and learning, the U.K. government's DfE funded a national rollout of NELI to state-maintained primary schools in England (publicly funded schools, similar to public elementary schools in the United States), prioritizing schools with relatively high levels of social deprivation. NELI was provided to 6,672 schools in 2020–2021 and an additional 4,442 schools in 2021–2022. The rollout received ethical approval from the University of Oxford's Department for Education Research Ethics Committee (reference: ED-CIA-21-005).

Participants

Across the 2 years of the rollout, 282,563 children in their first year of school (Reception class) were screened. Children with weak oral language skills were assigned to NELI based on their LanguageScreen standardized score (lowest 3–6 scores in the class). Schools permitted the use of LanguageScreen data for research purposes (https://media.oxedandassessment.com/assets/OxEdPrivStatement.pdf).

Our first RD analysis uses a data set that was provided to an independent evaluator of the rollout appointed by the EEF (National Foundation for Educational Research (Smith et al., 2023). This data set, henceforth the "Replication sample," includes children from a subset of schools from the second year (2021–2022) of the rollout. Only schools that had initially indicated willingness to take part in evaluation activities upon recruitment were invited to enroll in this evaluation. To be eligible for inclusion, schools needed to identify the children receiving NELI using LanguageScreen and repeat those assessments after delivering NELI. This first analysis serves as an important replication step, using a focused, rigorously collected sample. This enables us to verify that our analytic approach reproduces the findings obtained by an independent evaluator before extending the analysis to a larger sample, hence establishing the validity of our approach in a formally evaluated subset of the rollout.

Our second RD analysis applies the same analyses to a larger sample that includes data from all schools in both years of the national rollout that had a minimum of three children identified as allocated to NELI and three or more control children with pretest and posttest LanguageScreen scores in each participating class. This broader data set, referred to as the "Extension sample," captures the full real-world diversity of schools delivering NELI in the rollout. The purpose of this second analysis was to extend the findings from the initial replication sample, which had enrolled in an evaluation, to a more comprehensive sample of schools, thereby assessing the robustness and generalizability of the results across the full national implementation. Analyzing both data sets enables us to examine the consistency of effects across different implementation contexts, thereby providing a stronger and more policy-relevant assessment of program impact. The following components of the

rollout—assessment, intervention, training, and support—were structured to operationalize key principles of Implementation Science, drawing on the CFIR framework (Damschroder et al., 2009).

Assessment

LanguageScreen (Hulme et al., 2024) is an app-based language assessment running on a touchscreen tablet or smartphone. It assesses children's language skills using four subtests: expressive vocabulary, receptive vocabulary, listening comprehension, and sentence repetition. LanguageScreen takes approximately 10 min to administer, and all scoring is automated within the app.

The psychometric properties of LanguageScreen (Hulme et al., 2024) are excellent: It exhibits high test–retest reliability, a good fit to a Rasch measurement model, and minimal differential item functioning across key subgroups such as age, gender, and language background. This ensures the tool provides valid, reliable, and unbiased assessment across diverse populations. The tool's four subtests are added together to combine into a single total score based on a strong fit to the Rasch measurement model, supporting the concept of language ability as a unitary construct. From a psychometric perspective, this means that the total score is a sufficient statistic, providing a reliable and meaningful summary of a child's overall language skills. This makes LanguageScreen a robust tool for identifying language difficulties and monitoring progress following intervention. Results are provided as raw, percentile, and standard scores. Standard scores are age-adjusted standard scores, using 6-month age bands (e.g., 4;0–4;5 and 4;6–4;11 [years;months]), ensuring that each child's performance is interpreted relative to peers of a similar age. This standardization allows comparisons within and across classes of mixed ages and supports accurate identification of children with the greatest language needs. LanguageScreen scores reported throughout this study are standard scores.

Prior to the rollout, LanguageScreen had only been used as a research tool in the effectiveness trial of NELI (West et al., 2021). The rollout marked the first time the app was made directly available to schools. A new teacher-facing interface was developed to allow school staff to administer the assessment independently. Each school was provided with a secure online account through which they could access their pupils' results and monitor their progress. Schools setting up their accounts gave informed consent, via a privacy statement, to allow pupil data to be used for research purposes. The account displayed results both at the individual and class level, with children's scores ranked to help teachers easily identify those most in need of intervention. By enabling schools to manage their own assessment data, the system fostered local ownership, an important factor in promoting adoption and sustained use of LanguageScreen in the rollout.

An update introduced midway through the rollout enabled schools to record which children were receiving NELI; use of this feature was optional and could be applied retrospectively as well for children currently receiving NELI. Only schools providing LanguageScreen pretest and posttest scores and recording that children had received the NELI could be used in the current analyses.

Intervention

The NELI program focuses on improving children's vocabulary knowledge and their narrative, active listening, and speaking skills (https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/nuffield-early-language-intervention-neli). Vocabulary instruction follows a multicontextual approach, supporting deep learning through repeated exposure in varied contexts. Narrative work introduces children to story elements and the sequencing of events and encourages the use of expressive language and grammatical competence in using taught vocabulary in connected speech. Active listening work incorporates activities targeting auditory discrimination, sequencing, and rhyming, which are extended to include activities to enhance phonological awareness and reinforce letter-sound knowledge in the second half of the program. NELI is organized thematically: My body, Things we wear, People who help us, Growing, Journey, and Time. Children receive three small-group sessions in groups of between three and six children and two individual sessions a week for 20 weeks, delivered by trained teaching assistants. See Tables 1 and 2 for the session structure for both small-group and individual sessions and Supplemental Material S1 for two sample small-group sessions.

The small-group nature of NELI fosters a supportive environment to build children's confidence and encourage their active participation. A key feature of this environment is the use of Ted, a friendly mascot who acts as a nonjudgmental "listening partner" throughout the intervention. This helps to create a playful atmosphere, reduces performance anxiety, and motivates children to participate. The combination of personalized attention from the teaching assistant and the familiarity of Ted supports the development of positive relationships and a willingness to practice new language skills.

To support children in making progress and encourage sustained fidelity to the intervention, practitioners keep brief session records. Record forms are included in the program manual, with training provided to support practitioners in completing them. These records guide planning by prompting practitioners to set goals, track vocabulary learning, and monitor narrative development across linked sessions. This record-keeping prompts them to reflect on what the child achieved, note specific challenges or successes, and use these observations to inform

Table 1. Group session structure.

Small-group session	Purpose	Part 1 Timing (Weeks 1–10)	Part 2 Timing (Weeks 11–20)
Introduction	Give greeting, discuss the day of the week, revise listening rules, settle the children into session, play a listening game	3	2
Letter-sound	Introduce a new letter	—	3
Reinforcement	Reinforce vocabulary taught in the previous session, e.g., using flashcards	5	4
Vocabulary	Introduce new vocabulary, use flashcards	5	5
Narrative	Work to improve narrative skills, including sequencing and knowledge of story elements	10	9
Plenary	Sequence and revise session, award the best listener	2	2

Note. Timing is given in minutes.

future sessions, ensuring a responsive and developmentally informed approach to language support.

Training

The NELI program was supported by online training for practitioners, designed to promote consistent and high-fidelity implementation across a large number of participating schools. Training was delivered through invitation-only courses hosted on the FutureLearn platform. It was designed to be completed in approximately 10 hr based on feedback from schools regarding the amount of time they could reasonably allocate to training without adversely affecting completion rates. This time frame still ensured the training remained comprehensive, covering the content from the face-to-face format used in prior trials.

Three structured training courses were provided. Classroom teachers and teaching assistants in all participating classrooms completed an initial course that covered oral language development and strategies to support children's language development and gave an overview of the NELI program. Including classroom teachers in this first course was intended to foster buy-in and equip them with knowledge to actively support the teaching assistant's delivery, ensuring they were engaged partners in the intervention rather than potential barriers to its implementation. A second course, delivered to the teaching assistants only, provided in-depth information about NELI and how to deliver it, including scripted session delivery, managing small-group and individual work and tracking children's progress. A final course halfway through delivery of the intervention trained practitioners in the basics of teaching phonics, with an emphasis on supporting the development of letter-sound knowledge and phoneme awareness through activities embedded in the final 10 weeks of NELI.

An important feature of the online training model was the deliberate fostering of a NELI community of practice, through interactive components embedded within every step of the online training. Practitioners were encouraged to comment, ask questions, and reflect, supporting shared learning and the integration of program practices into everyday routines.

To complement the training, a dedicated team of NELI mentors provided on-going real-time support. Mentors actively monitored discussions on the training platform, responding to queries, clarifying course content, and supporting implementation. A centralized guidance document, "The Mentor Handbook," was used across the team to ensure that responses were aligned and evidence-based. This shared resource included agreed answers to frequently asked questions, guidance on nuanced delivery decisions, and procedures for flagging more complex queries for expert review. In this way, the mentoring system ensured that support was both immediate and reliable, while still allowing for escalation when more considered or context-specific input was needed.

Mentors also used FutureLearn's monitoring functionality to oversee participation at scale, tracking key indicators such as invitation acceptance, course progression, completion status, and end-of-course assessment performance. These data

Table 2. Individual session structure.

Individual session	Purpose	Timing
Introduction	Give greeting, revise listening rules if needed, settle the child into the session	2
Vocabulary	Revise vocabulary the child has found difficult or focus on advanced uses	5
Narrative	Scaffold the child's narrative development using a picture story sequence or a personal event narrative across the week's two sessions	5
Plenary	Revise the session and give a reward sticker	3

Note. Timing is given in minutes.

enabled a small mentor team to identify schools or individuals with stalled engagement and provide timely, targeted support to ensure practitioners were well prepared for program delivery.

In the 2 years of the rollout, 18,506 teachers and teaching assistants successfully completed the first joint training course, with 13,310 teaching assistants going on to complete their second course (please note that fewer teaching assistants completed the third course midway through delivery; n = 7,249). Qualitative feedback from optional end-of-training surveys indicated that the training was well received, with trainees reporting increased knowledge about language development, enhanced motivation to support children's language learning, and greater confidence in delivering NELI. This suggests that the training effectively prepared educators to deliver the intervention.

Support

In addition to formal training, practitioners had access to the NELI Delivery Support Hub, an online resource designed to offer flexible, needs-led support throughout program delivery. The Hub served as a central repository for additional content, including a video library of model sessions, exemplar record forms, lesson plan suggestions tailored to different levels of language ability (the "Teddy Levels"), and practical advice on running both group and individual sessions. The "See NELI in Action" video library was particularly valuable for new practitioners, featuring real-life videos of NELI sessions delivered by NELI program experts, along with associated session records to model effective delivery and illustrate best-practice approaches to progress tracking.

The Hub further developed the "community of practice" approach, encouraging social learning and peer-to-peer support through discussion threads, comment sections, and mentor moderation. Practitioners were invited to share questions, tips, and success stories and to engage with the contributions of others. A dedicated "SOS Help Centre" enabled the mentor team to respond to any queries that were not already addressed in the resource sections. This multifaceted support structure, combining formal training, practical examples, self-service resources, and peer interaction, was designed to ensure that all practitioners were equipped not only to begin delivering the program confidently but also to maintain high fidelity and responsiveness throughout the implementation process.

Finally, in addition to the online support provided to NELI practitioners in the Hub, schools had access to a dedicated NELI call center for further assistance. This central helpdesk could be contacted via live chat, e-mail, or telephone, providing an additional layer of responsive support. The call center handled a wide range of queries, from technical issues accessing the training platform to questions about program delivery, ensuring that schools could receive timely and practical help when needed.

Analysis

An RD design provides a robust way of evaluating the effectiveness of an intervention when a randomized control group is not available (Cattaneo & Titiunik, 2022; Waddington et al., 2023). A classic RD analysis requires that participants receive a pre-intervention or "pretest" score and that an intervention be assigned on the basis of this score. Assignment of the intervention is determined by a participant's pretest score relative to a prespecified cutoff: Participants whose scores fall on one side of the cutoff are assigned to the intervention, while those on the other side of the cutoff are not. The analysis then compares the outcome between participants whose pretest scores are just above the cutoff and participants whose pretest scores are just below it (Trochim & Donnelly, 2001). The fundamental assumption is that participants whose scores are just above the cutoff are similar in observed characteristics to participants whose scores are just below the cutoff except in the assignment of the intervention, and thus, the comparison between their outcomes can be a reliable measure of the true causal effect of the intervention. Intuitively, this assumption is likely to be satisfied if participants have no control over the specific value of their score, so that if their score falls just above the cutoff, it is plausible to imagine that it could have fallen just below it. The main strategy is therefore to restrict the analysis only to participants whose scores are very near the cutoff, where the assumption of comparability is most plausible. Regression functions are calculated for each group, and a discontinuity between these functions at the assignment cutoff indicates a treatment effect (Cappelleri & Trochim, 2003). On a graph, an RD, therefore, appears as two lines of best fit (one for each group), one to the left and one to the right of the cutoff; a visible "jump" or gap between these lines at the cutoff indicates the size of the treatment effect.

The most common approach for RD analysis uses a continuous assignment score and is based on assumptions of continuity of regression functions at the cutoff. In our study, this continuity-based approach is not applicable, because LanguageScreen scores take only integer values, leading to multiple children having the same value as their score. Instead, we used a local randomization RD approach, which is based on the assumption that close to the cutoff, treated and control groups resemble a randomized experiment (Cattaneo et al., 2020, 2024).

For our analyses, we used LanguageScreen standard scores for both the baseline RD score and the posttest RD outcome. A major advantage of using LanguageScreen in this context is that it produces standard scores—norm-referenced scores with a fixed mean and standard deviation. This means a given score change represents the same amount of improvement across the entire distribution, enabling meaningful comparisons between children with different baseline language abilities. This is particularly important in

our RD analysis, as it allows gains to be interpreted consistently regardless of where children begin. Our RD analyses involve two modifications to a classic RD analysis necessitated by complexities resulting from the assignment guidance given to schools and consequent disparities in baseline LanguageScreen scores between comparison groups.

Establishing the Cutoff for Assignment to Intervention

Schools delivering NELI in the national rollout were instructed by the DfE to select between three and six children in each classroom with the lowest baseline (prestest) LanguageScreen scores to receive NELI. The allocation rule was, therefore, based on a variable range of the ranks of this baseline LanguageScreen score, so the cutoff is not directly known. In other words, because schools had latitude to choose a different number of children with the lowest scores in each class, there was not a single fixed score that determined who got the intervention. To deal with this issue, we followed the procedure of Smith et al. (2023) in their independent evaluation of some of the data reported here. The cutoff in each classroom was imputed as the LanguageScreen score that minimized discrepancies between treatment assignment and actual treatment received. Specifically, for each class, we identified the cutoff in LanguageScreen pretest score that resulted in the greatest number of children who received NELI being correctly assigned to the intervention condition and the greatest number who did not receive NELI being correctly assigned to the control condition. A normalized score was then created by subtracting the imputed cutoff from the pretest LanguageScreen score. This rescaling meant that the cutoff in every class was set to zero. The normalized score was then multiplied by −1 so that children with scores greater than or equal to zero were those assigned to NELI. Normalizing the scores in this way placed all classrooms onto a common scale. As a result, children's positions relative to their own class-specific cutoff could now be directly compared across all classes, enabling us to pool the data for analysis using a single, consistent cutoff point. We refer to this score as the normalized pretest LanguageScreen score, or simply as the pretest normalized score.

Selecting the Outcome Measure

The goal of an RD analysis is to compare participants just above the cutoff with participants just below it. However, as discussed above, this study departs from the ideal RD design, because schools were permitted to choose between three to six children to treat. This meant that schools quite rationally selected clusters of low-performing children whose LanguageScreen scores were considerably lower than the rest of the children in the classroom. For example, in one school, teachers selected three children for NELI with standard scores of 73, 77, and 85, while the next children in rank order, scoring 95, 96, 97, 97, 97, and 99, were not selected, resulting in a gap of 10 points between the last child assigned to NELI (85) and the first child not assigned to NELI (95). As a result, even when restricting the analysis to children just assigned to NELI and just not assigned to NELI—that is, children with pretest scores at or just above the imputed cutoff and children with pretest scores just below the imputed cutoff—children assigned to NELI began with systematically poorer language skills than their peers, and in some classrooms, this disparity was pronounced. A comparison of barely assigned-to-treatment and barely assigned-to-control children using solely the posttest outcome would therefore be misleading, as it would not take into account the systematic differences in baseline LanguageScreen scores between the two groups. For this reason, the outcome used in the RD analyses in the current article is the change between baseline and posttest score on LanguageScreen for each child. Using change scores as the outcome allows us to account for differences in initial language skills by measuring each child's progress during the intervention. This approach ensures that any interpretation of posttest outcomes considers where children started, avoiding misleading conclusions about the intervention's effectiveness; in fact, equivalent posttest scores would demonstrate that NELI children have effectively closed their initial language gap.

For implementation of a local randomization RD design analysis, we first restrict the analysis to children whose normalized pretest LanguageScreen scores fall within a small neighborhood or "window" around the imputed cutoff, and within this window, we compare the outcomes of children whose normalized pretest scores are above the cutoff to the outcomes of children whose normalized pretest scores are below the cutoff. These so-called "intention-to-treat" analyses capture the effect of assigning children to NELI based on the imputed cutoff. This differs from the effect of actually giving children the intervention, which cannot be calculated because there is imperfect compliance: Some children whose normalized pretest scores are above the cutoff do not receive NELI, while some children whose normalized pretest scores are below the cutoff receive NELI anyway. However, given how the cutoff is imputed, compliance is very high. In our windows of analysis, only about 5% children with scores below the cutoff receive NELI, and roughly 85% of the children with scores above the cutoff receive NELI. Supplemental Material S2 contains the R code used for all analyses.

Results

Replication of Previous Findings

The data used in this first RD analysis contained all children with nonmissing baseline LanguageScreen scores from schools opting in to the independent evaluation of

the rollout (n = 548 schools, 19,212 children). This included 7,347 children with missing posttest outcomes who were dropped from the analysis, leaving 11,865 children (n = 466 schools) with baseline and posttest scores (see Replication Sample columns in Table 3 for descriptive information on this sample).

The choice of window must balance the number of observations with the comparability of the observations included: The smaller the window, the more likely children with scores above the cutoff will be similar to children with scores below the cutoff, but the fewer children will be included in analysis. We report results for the window [−3, 2], the third smallest window around the imputed cutoff. In this window, children are balanced in terms of age and gender (difference in age in months is −0.21, p = .29; difference in share of boys is 0.03, p = .24), although they are significantly different in whether English is an additional language (difference in EAL share is 0.06, p < .01). This difference in EAL share is unsurprising because assigned-to-treatment children have lower pretest LanguageScreen scores than assigned-to-control children by construction, and children with EAL status are more likely to have lower LanguageScreen scores. Robustness analyses in the smallest possible window [−1, 0] yield similar patterns, which supports the integrity of the results because this narrow window includes only children whose scores are closest to the cutoff, making the assigned-to-treatment and assigned-to-control groups highly comparable and strengthening causal inference despite the smaller sample size (see Table 4 for effect sizes in all windows).

Children assigned to NELI (i.e., those at or above the cutoff) improved by 3.27 standard score points more than children below the cutoff who had not received the intervention (NELI n = 1,115, control n = 541; p < .01; Hedges's g = 0.36, 95% CI [0.26, 0.46]). The mean improvement between posttest and pretest LanguageScreen standard scores for NELI children at or above the cutoff is 12.05 (baseline mean = 87.70, posttest mean = 99.75), while the mean improvement among children below the cutoff who have not received NELI is 8.78 (baseline mean = 92.85, posttest mean = 101.64). In other words, children assigned to NELI (according to the imputed cutoff) have substantively closed the gap between themselves and their peers in LanguageScreen scores that existed before the intervention. Figures 1a and 1b show the discontinuity for the functions relating pretest to posttest scores for the two groups in this analysis. Figure 1a presents the full RD plot, which shows the change between pretest and posttest LanguageScreen scores against the normalized pretest LanguageScreen score, displaying all children in the Replication sample, with the cutoff indicated as a vertical line. There is a distinct discontinuity at the cutoff, indicating that children just assigned to NELI have a higher average improvement in LanguageScreen scores between pretest and posttest than children who just missed assignment. Figure 1b zooms in on the local region around the cutoff by restricting the sample to observations within the [−3, 2] window, providing a more local visualization of the treatment effect. In both figures, each dot represents the average change in LanguageScreen scores for small groups (bins) of children who have similar normalized pretest scores. The fit lines incorporate both linear and quadratic terms to capture any curvature in the relationship on either side of the cutoff, rather than assuming a strictly linear trend.

Extension to a Larger Sample

Our second analysis includes all schools in both years of the national rollout that provided pretest and posttest LanguageScreen scores for at least three identified NELI and three non-NELI children within each classroom. This resulted in a total sample size of 19,936 children, an additional 8,071 children from a further 301 schools compared to the previous analysis (see Extension sample columns in Table 3 for descriptive information on this sample).

The analysis used the same procedures as the previous analysis. Children assigned to NELI improved by 3.44 standard score points more than children below the cutoff

Table 3. Descriptive information on all children included in replication and extension regression discontinuity analyses.

Variable	Replication sample, *M* (*SD*)			Extension sample, *M* (*SD*)		
	All children (n = 11,865)	NELI children (n = 2,921)	Control children (n = 8,944)	All children (n = 19,936)	NELI children (n = 5,084)	Control children (n = 14,852)
Age at T1	57.55 (3.74)	57.59 (4.02)	57.54 (3.64)	57.53 (3.79)	57.50 (4.02)	57.54 (3.71)
Gender	0.51 (0.50)	0.56 (0.50)	0.50 (0.50)	0.51 (0.50)	0.56 (0.50)	0.50 (0.50)
EAL	0.20 (0.40)	0.36 (0.48)	0.15 (0.36)	0.21 (0.41)	0.38 (0.48)	0.15 (0.36)
T1 LanguageScreen	100.30 (15.56)	83.01 (10.41)	105.95 (12.50)	99.23 (15.42)	82.46 (9.97)	104.97 (12.48)
T2 LanguageScreen	107.20 (14.57)	95.06 (13.82)	111.16 (12.46)	106.07 (14.45)	94.51 (13.62)	110.02 (12.47)

Note. "NELI children" refers to children assigned to NELI according to the imputed cutoff, and "control children" refers to children not assigned to NELI according to the imputed cutoff, regardless of whether they received NELI or not. T1 refers to pretest, and T2 refers to posttest. LanguageScreen are standardized scores. NELI = Nuffield Early Language Intervention; EAL = English as an additional language.

Table 4. Differences in delta standardized LanguageScreen scores across regression discontinuity analysis windows by cutoff status (Nuffield Early Language Intervention [NELI] and control).

Window	NELI mean Δ	Control mean Δ	Diff	95% CI	*p*	NELI *n*	Control *n*	Hedges's *g*	Hedges's *g* 95% CI
Replication analysis									
[–1, 0]	12.01	7.84	4.17	[2.75, 5.61]	> .001	760	144	0.54	[0.36, 0.72]
[–2, 1]	12.13	9.19	2.94	[1.84, 4.04]	> .001	926	344	0.34	[0.21, 0.46]
[–3, 2]	12.05	8.78	3.27	[2.33, 4.21]	> .001	1,115	541	0.36	[0.26, 0.46]
Extension analysis									
[–1, 0]	11.68	8.53	3.15	[2.09, 4.22]	> .001	1,277	264	0.41	[0.28, 0.54]
[–2, 1]	11.86	8.69	3.16	[2.37, 3.96]	> .001	1,583	624	0.38	[0.29, 0.47]
[–3, 2]	11.84	8.40	3.44	[2.76, 4.12]	> .001	1,920	981	0.40	[0.32, 0.47]

Note. Results use imputed cutoff. NELI mean Δ and Control mean Δ = difference between pre and post LanguageScreen standard scores; Diff = difference in mean delta between the NELI and control groups; CI = confidence interval.

who were not assigned to the program (NELI n = 1920, control n = 981; $p < .01$; Hedges's g = .40, 95% CI [.32, .47]). The mean improvement between posttest and pretest LanguageScreen standard scores for children assigned to NELI according to the imputed cutoff was 11.84 LanguageScreen standard scores (pretest mean = 87.24, posttest mean = 99.08), while the improvement for children below the cutoff was 8.40 (baseline mean = 91.69, posttest mean = 100.09). Figures 2a and 2b show the discontinuity for the functions relating pretest to changes between pre- and posttest scores for the two groups in this analysis.

Converging Evidence for the Effectiveness of the NELI Program

The analyses above are not standard RD analyses and are better viewed as "RD-like" because schools did not follow a rigid, prespecified rule for assigning children to the intervention. We addressed this complication by analyzing the change in LanguageScreen standard scores between pretest and posttest as our outcome. The results showed that children who had received NELI made greater gains in LanguageScreen standard scores than the children who had just missed assignment to the program.

A causal interpretation of this improvement in LanguageScreen standard scores requires the assumption that, in the absence of the NELI, the average improvement for children with scores just above the cutoff would have been the same as for children with scores just below it. This is a strong assumption that is difficult to verify. If regression to the mean were operating, it would violate this assumption, since retested children with lower scores (i.e., those assigned to NELI) would be expected to show greater regression toward the mean (greater increases in their scores) than children with higher scores who were not assigned to intervention.

We conducted an additional analysis to address the threat of regression to the mean (and also address any concerns regarding the cutoff imputation) by conducting analyses using nearest-neighbor matching (Stuart, 2010), which make minimal parametric assumptions and are designed to provide unbiased estimates of the average treatment effect. We performed nearest-neighbor matching of children who received NELI to children who did not receive it, using four pretest measures: LanguageScreen score, age, EAL status (whether children came from a home where a language other than English was the main language spoken), and gender (see Table 5 for descriptive information on the children included in this analysis). In this analysis, for every child who actually received NELI, we find a child who did not receive NELI but is as similar as possible (or "nearest") in terms of age, EAL status, and gender and exactly matched on pretest LanguageScreen score.

We matched exactly on the pretest LanguageScreen score and implemented the analysis with the matching R package (Sekhon, 2011). The results of this analysis were consistent with those from the RD analysis above. LanguageScreen standard scores at posttest for children receiving NELI were 3.92 points higher than matched control children (n: NELI = 4,802, control = 4,802; mean NELI posttest LanguageScreen standard score = 96.88; mean control posttest score = 92.96; SE = .26; $p < .01$; Hedges's g = .32, 95% CI [.28, .36]). Critically, nearest-neighbor matching eliminates concerns about regression to the mean because each treated child is matched with a control child who has the same pretest LanguageScreen score. These analyses, therefore, provide converging evidence for the effectiveness of the NELI program in the national rollout and indicate that the results from our RD analyses are unlikely to have been caused by regression to the mean.

Discussion

The analyses reported here demonstrate that the NELI program (a targeted language intervention for children with

Figure 1. (a) Replication RD analysis: Effect of crossing imputed cutoff on increases in LanguageScreen scores, using all observations (n = 11,865 children). Dots are sample means of the outcome within bins of the normalized pretest score. Red line is second-order global polynomial fit, separately calculated for observations above and below the cutoff. (b) Replication RD analysis: effect of crossing imputed cutoff on increases in LanguageScreen scores, using only observations in the [−3, 2] window (n = 1,656 children). Dots are sample means of the outcome within bins of the normalized pretest score. Red line is overall mean (zero-order polynomial fit), separately calculated for observations above and below the cutoff. RD = regression discontinuity.

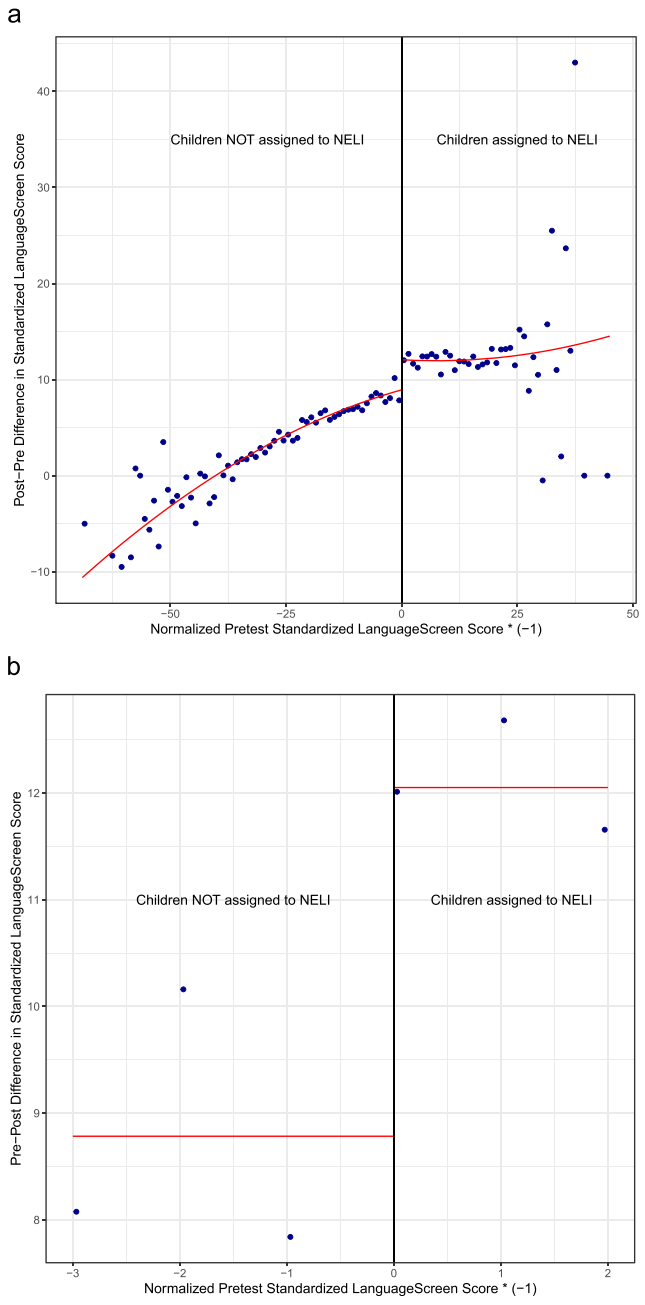

Figure 2. (a) Extension RD analysis: Effect of crossing imputed cutoff on increases in LanguageScreen scores, using all observations (n = 19,936 children). Dots are sample means of the outcome within bins of the normalized pretest score. Red line is second-order global polynomial fit, separately calculated for observations above and below the cutoff. (B) Extension RD analysis: effect of crossing imputed cutoff on increases in LanguageScreen scores, using only observations in the [−3, 2] window (n = 1,656 children). Dots are sample means of the outcome within bins of the normalized pretest score. Red line is overall mean (zero-order polynomial fit), separately calculated for observations above and below the cutoff. RD = regression discontinuity.

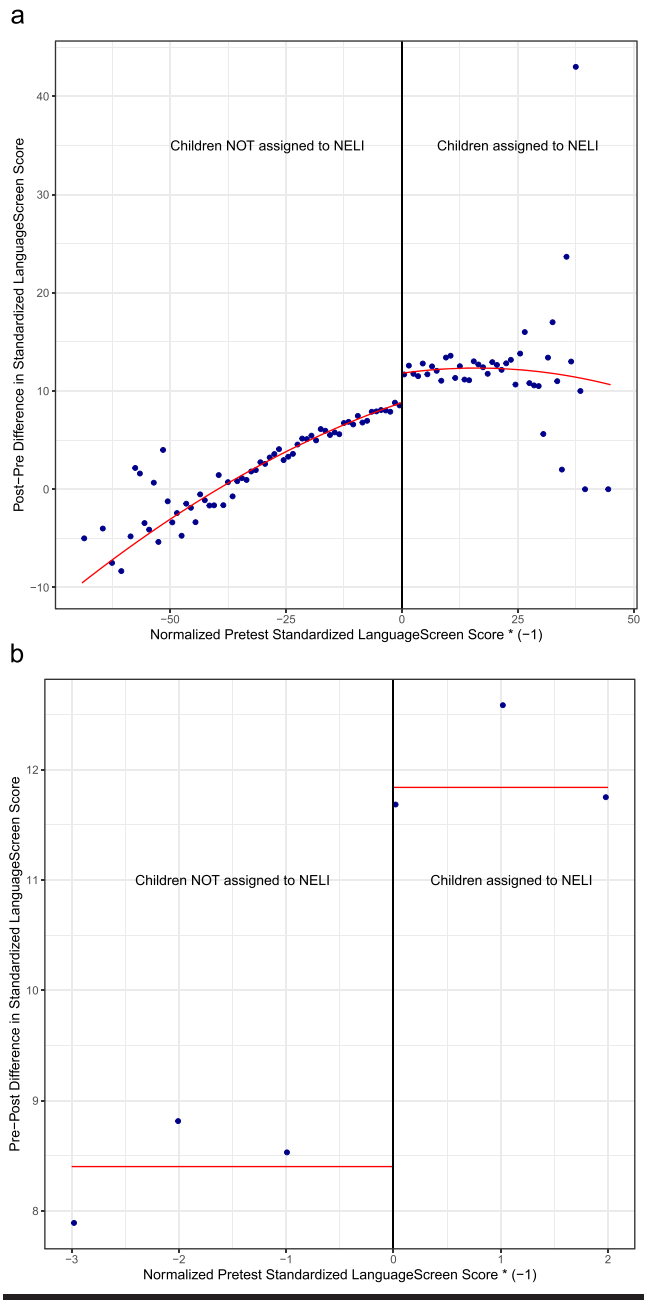

Table 5. Descriptive information of children in sensitivity analysis (nearest-neighbor matching).

		Before matching		After matching	
	All children	NELI children	Control children	NELI children	Control children
Variable	*M* (*SD*)	*M* (*SD*)	*p*	*M* (*SD*)	*p*
Age at T1	57.45 (4.03)	57.58 (3.64)	.13	57.34 (3.60)	> .001
Gender	0.55 (0.50)	0.50 (0.50)	> .001	0.55 (0.50)	.16
EAL	0.36 (0.48)	0.15 (0.36)	> .001	0.36 (0.48)	1.00
T1 LanguageScreen	84.84 (10.79)	104.98 (13.64)	> .001	84.84 (10.79)	1.00
T2 LanguageScreen	97.12 (13.46)	110.25 (13.48)	> .001	93.85 (12.89)	> .001

Note. Results use the Replication sample. NELI children refers to children who received NELI, and control children refers to children who did not receive NELI. The *p* values reported correspond to a test of the null hypothesis that the mean among NELI children is the same as the mean among control children. T1 refers to pretest, and T2 refers to posttest. NELI = Nuffield Early Language Intervention; EAL = English as an additional language

language weaknesses) produced educationally meaningful improvements in language skills in a national rollout to schools in England. The results are striking and indicate that the program delivered in the rollout produced educationally meaningful improvements in language skills, with effect sizes that were equivalent to or slightly larger than those found in a previous large-scale RCT (West et al., 2021).

The measure used in the analyses was LanguageScreen (Hulme et al., 2024), a broad measure of receptive and expressive language skills with high reliability and validity. Schools used it to allocate children to NELI and evaluate progress. The first RD analysis was conducted with data from schools opting in to an independent evaluation of the national rollout of NELI. Analyses of these data by the independent evaluator (Smith et al., 2023) showed significant improvements in LanguageScreen scores as a result of NELI (d = 0.30). Our reanalyses of these data included data from roughly 1,000 more children and confirmed this pattern, although we found a slightly larger improvement in children's LanguageScreen scores (effect size Hedges's g = .34), which likely reflects differences in the sample and minor differences in analytic choices. Our second analysis combined data from the independent evaluation study with data from other schools and another phase of the rollout to give a much larger sample of almost 20,000 children. This analysis showed a larger improvement in LanguageScreen scores (Hedges's g = .40) than found for the independent evaluation sample alone. This finding was also replicated using a nearest-neighbor matching design that addressed the lack of comparability of barely treated and barely control children in the RD design that was inadvertently introduced by the flexible treatment allocation guidelines given to schools.

There has been widespread concern in medicine (Damschroder et al., 2009) and education (Snowling et al., 2022) that evidence-based interventions often do not scale effectively. Obstacles to effective scaling have been identified that operate at many different levels (Damschroder et al., 2009), including properties of the intervention (its quality and evidence base), the recipients (do patients accept or comply with an intervention), and the context of delivery (organizational leadership's backing for an intervention). The results reported here, from a very large sample in a nationwide rollout, demonstrate that an evidence-based intervention such as NELI, when well implemented, can remain effective at scale. This stands in contrast to concerns about the lengthy delays in translating research into practice—widely cited as averaging 17 years (Balas & Boren, 2000), though later analyses, including Morris et al. (2011), highlight the lack of standardized definitions and metrics to accurately capture such lags. The trajectory of NELI, from its first trial published in 2008 to nationwide implementation more than a decade later, underscores the persistence of lengthy research-to-practice timelines, even for interventions with strong evidence and policy relevance.

One reason for such an extended timeline is that university departments and research teams, while well suited for developing and testing interventions, typically lack the infrastructure, continuity, and operational capacity required for large-scale delivery over time. In the case of NELI, these challenges were mitigated by consistent leadership guiding the program from its early trials at the University of York (Bowyer-Crane et al., 2008) to its large-scale effectiveness trial under more naturalistic conditions at the University of Oxford (West et al., 2021). The first year of the national rollout was led by the Oxford research team. However, to support sustainable delivery of evidence-based tools in education, including the national rollout of NELI, a University of Oxford spinout company, OxEd and Assessment Ltd, was established. This organization, staffed largely by members of the Oxford research team, led the second year of the rollout with the goal of enabling broad and lasting impact in schools.

Several factors identified as important in studies of implementation science may help to explain the success demonstrated here for the NELI program in this rollout.

These align closely with the five domains in Damschroder et al.'s (2009) CFIR, although our focus here is on their practical implications for schools. First is the quality of the intervention. NELI is an intervention built on sound pedagogical principles (Beck et al., 2013; Davies, et al., 2004), in which those delivering the intervention learn to use "scaffolding" to help children develop core aspects of oral language, including vocabulary knowledge and narrative and active listening skills. The program was delivered unchanged from the version used in the effectiveness trial, preserving its theoretical integrity while presenting materials and activities in a format accessible to nonspecialists, which supported consistent, high-quality delivery across diverse school contexts (CFIR: Intervention Characteristics).

Second, the development of online training and support was designed to upskill teachers and teaching assistants to deliver the program as intended. From the start, this training and support focused on increasing knowledge and changing beliefs about language development and teaching, with emphasis placed on addressing any perceived difficulty of implementation, reflected by program intensity and duration and potential disruption to school routine. We believe that the high-quality online training, which made school staff knowledgeable about and committed to the program, was one critical component of the success of the rollout.

The training and support infrastructure for NELI was intentionally designed to address multiple domains of the CFIR framework. For example, the inclusion of role-specific modules, demonstration videos, and opportunities for peer interaction built practitioners' knowledge, strengthened delivery confidence, and created a supportive professional community (CFIR: Inner Setting, Characteristics of Individuals). Mentoring structures and accessible help channels enabled timely resolution of challenges and sustained engagement (CFIR: Process). Together, these elements created a supportive implementation climate that promoted fidelity and responsiveness. Third, the use of LanguageScreen before and after delivery of NELI provided reliable baseline and postdelivery child-level measures, which could be used to evaluate the impact of the rollout on language skills. Providing teachers with LanguageScreen via a dedicated school account gave them full control over identifying children's needs and tracking progress, minimizing extra administrative burden and making assessment a seamless part of their practice. This approach supported decision making at all levels, from classrooms to senior leadership and policymakers, while also enabling robust evaluation of the program (CFIR: Process, Outer Setting).

Building on this, the current article highlights an important methodological consideration: Evaluating programs at scale requires rigorous analytic approaches that leverage such reliable data. RD designs and nearest-neighbor matching models potentially have great promise in evaluating the effectiveness of programs that have evidence from trials and are now being implemented at scale (e.g., Ludwig & Miller, 2007). Such promise depends on having reliable and valid measures of improvement. In the current study, LanguageScreen provided such a measure.

Next Steps

The first trial evaluating NELI was published over 15 years ago (Bowyer-Crane et al., 2008), supporting the previous conclusion from implementation science that it typically takes around 17 years for just 14% of original research to be translated into effective practice (Balas & Boren, 2000). For NELI, government funding directed toward COVID-19 recovery helped to bridge the "last mile" to implementation at a national scale. Finding ways to speed the journey of other promising educational interventions from trials to delivery at scale without sacrificing quality is an important future challenge for researchers and policymakers alike.

Looking ahead, NELI is well suited to being embedded within an MTSS in education, in which interventions are matched to children's level of need and adjusted based on response to intervention. The integration of LanguageScreen into the program ensures that screening for oral language difficulties is placed directly in the hands of teachers. This enables early identification and timely support in the first year of school, while also generating reliable data that can be shared with speech and language professionals or used to escalate support where needed. For children with developmental language disorder, in particular, who typically require longer term and more specialized intervention, this framework facilitates timely referral and tailored support beyond Tier 2 interventions.

Limitations

The current analyses include measures from nearly 20,000 children in classes delivering NELI, giving us high statistical power. The analyses only included data from schools that completed online reports of which children were receiving NELI, and the ability to record this information was only introduced midway through the 2 years of the rollout. We cannot be sure, however, that the schools whose data were analyzed here are representative of the rest of the schools that received the NELI program as part of the rollout.

Conclusions

These results have important implications for practice and policy in education. We have shown that the NELI program can be effective in improving children's language

skills when rolled out nationally. Language difficulties place children at high risk of a range of adverse outcomes including educational failure, social and emotional difficulties, and reduced employment prospects (Clegg et al., 2005). Early interventions, such as the NELI program, can be implemented at scale at relatively low cost and have great promise for improving well-being and reducing the adverse educational, mental health, and economic consequences associated with language difficulties.

Data Availability Statement

Schools in this study permitted the use of LanguageScreen data for research purposes (https://media.oxedandassessment.com/assets/OxEdPrivStatement.pdf). Individual participant data are, therefore, not available for sharing due to legal restrictions. Analysis scripts are available as supplemental materials.

Acknowledgments

The rollout was funded by two grants from the Educational Endowment Foundation (EPR01500 and EPR01560) to Charles Hulme and Gillian West. Rocío Titiunik gratefully acknowledges financial support from the National Science Foundation through Grant SES-2241575. General purpose RD software is available at https://rdpackages.github.io/.

References

Balas, E. A., & Boren, S. A. (2000). Managing clinical knowledge for health care improvement. *Yearbook of Medical Informatics, 9*(1), 65–70. https://doi.org/10.1055/s-0038-1637943

Beck, I. L., McKeown, M. G., & Kucan, L. (2013). *Bringing words to life: Robust vocabulary instruction* (2nd ed.). Guilford Press. https://www.guilford.com/books/Bringing-Words-to-Life/Beck-McKeown-Kucan/9781462508167

Bloom, H. S. (2012). Modern regression discontinuity analysis. *Journal of Research on Educational Effectiveness, 5*(1), 43–82. https://doi.org/10.1080/19345747.2011.578707

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

Cappelleri, J. C., & Trochim, W. M. (2003). Cutoff designs. In S.-C. Chow (Ed.), *Encyclopedia of biopharmaceutical statistics* (2nd ed., pp. 263–269). Marcel Dekker. https://www.academia.edu/84883545/Cutoff_Designs

Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2020). *A practical introduction to regression discontinuity designs: Foundations*. Cambridge University Press. https://doi.org/10.1017/9781108684606

Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2024). *A practical introduction to regression discontinuity designs: Extensions*. Cambridge University Press. https://doi.org/10.1017/9781009441896

Cattaneo, M. D., & Titiunik, R. (2022). Regression discontinuity designs. *Annual Review of Economics, 14*(1), 821–851. https://doi.org/10.1146/annurev-economics-051520-021409

Clegg, J., Hollis, C., Mawhood, L., & Rutter, M. (2005). Developmental language disorders—A follow-up in later adult life. Cognitive, language and psychosocial outcomes. *Journal of Child Psychology and Psychiatry, 46*(2), 128–149. https://doi.org/10.1111/j.1469-7610.2004.00342.x

Damschroder, L. J., Aron, D. C., Keith, R. E., Kirsh, S. R., Alexander, J. A., & Lowery, J. C. (2009). Fostering implementation of health services research findings into practice: A consolidated framework for advancing implementation science. *Implementation Science, 4*(1), 1–15. https://doi.org/10.1186/1748-5908-4-50

Damschroder, L. J., Reardon, C. M., Widerquist, M. A. O., & Lowery, J. (2022). The updated Consolidated Framework for Implementation Research based on user feedback. *Implementation Science, 17*(1), Article 75. https://doi.org/10.1186/s13012-022-01245-0

Davies, P., Shanks, B., & Davies, K. (2004). Improving narrative skills in young children with delayed language development. *Educational Review, 56*(3), 271–286. https://doi.org/10.1080/0013191042000201181

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

Fricke, S., Bowyer-Crane, C., Snowling, M. J., & Hulme, C. (2018). *The Nuffield Early Language Intervention*. Oxford University Press.

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

Gaias, L. M., Cook, C. R., Brewer, S. K., Bruns, E. J., & Lyon, A. R. (2023). Addressing the "last mile" problem in educational research: Educational researchers' interest, knowledge, and use of implementation science constructs. *Educational Research and Evaluation, 28*(7–8), 205–233. https://doi.org/10.1080/13803611.2023.2285440

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., Hansen, J. H. L., & Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology, 26*(2), 248–265. https://doi.org/10.1044/2016_AJSLP-15-0169

Gleason, P. M., Resch, A. M., & Berk, J. A. (2012). *Replicating experimental impact estimates using a regression discontinuity approach* (NCEE Reference Report 2012-4025). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children.* Brookes.

Hessel, A. K., & Strand, S. (2023). Proficiency in English is a better predictor of educational achievement than English as an additional language (EAL). *Educational Review, 75*(4), 763–786. https://doi.org/10.1080/00131911.2021.1949266

Hjetland, H. N., Brinchmann, E. I., Scherer, R., Hulme, C., & Melby-Lervag, M. (2020). Preschool pathways to reading comprehension:

A systematic meta-analytic review. *Educational Research Review, 30,* Article 100323. https://doi.org/10.1016/j.edurev.2020.100323

Hulme, C., McGrane, J., Duta, M., West, G., Cripps, D., Dasgupta, A., Hearne, S., Gardner, R., & Snowling, M. (2024). LanguageScreen: The development, validation, and standardization of an automated language assessment app. *Language, Speech, and Hearing Services in Schools, 55*(3), 904–917. https://doi.org/10.1044/2024_LSHSS-24-00004

Hulme, C., Nash, H. M., Gooch, D., Lervåg, A., & Snowling, M. J. (2015). The foundations of literacy development in children at familial risk of dyslexia. *Psychological Science, 26*(12), 1877–1886. https://doi.org/10.1177/0956797615603702

Hulme, C., Snowling, M. J., West, G., Lervåg, A., & Melby-Lervåg, M. (2020). Children's language skills can be improved: Lessons from psychological science for educational policy. *Current Directions in Psychological Science, 29*(4), 372–377. https://doi.org/10.1177/0963721420923684

Komesidou, R., & Hogan, T. P. (2023). A generic implementation framework for school-based research and practice. *Language, Speech, and Hearing Services in Schools, 54*(4), 1165–1172. https://doi.org/10.1044/2023_LSHSS-22-00171

Lester, N., Twomey, K. E., & Theakston, A. (2025). What difficulties do children learning English in addition to another language experience with English oral language? A systematic review. *The Language Learning Journal, 53*(2), 264–285. https://doi.org/10.1080/09571736.2024.2369291

Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *The Quarterly Journal of Economics, 122*(1), 159–208. https://doi.org/10.1162/qjec.122.1.159

Morris, Z. S., Wooding, S., & Grant, J. (2011). The answer is 17 years, what is the question: Understanding time lags in translational research. *Journal of the Royal Society of Medicine, 104*(12), 510–520. https://doi.org/10.1258/jrsm.2011.110180

Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., Vamvakas, G., & Pickles, A. (2016). The impact of nonverbal ability on prevalence and clinical presentation of language disorder: Evidence from a population study. *Journal of Child Psychology and Psychiatry, 57*(11), 1247–1257. https://doi.org/10.1111/jcpp.12573

Norbury, C. F., Griffiths, S., Vamvakas, G., Baird, G., Charman, T., Simonoff, E., & Pickles, A. (2021). *Socioeconomic disadvantage is associated with prevalence of developmental language disorders, but not rate of language or literacy growth in children from 4 to 11 years: Evidence from the Surrey Communication and Language in Education Study (SCALES)*. Preprints With The Lancet. https://doi.org/10.2139/ssrn.3814832

Pace, A., Luo, R., Hirsh-Pasek, K., & Golinkoff, R. M. (2017). Identifying pathways between socioeconomic status and language development. *Annual Review of Linguistics, 3*(1), 285–308. https://doi.org/10.1146/annurev-linguistics-011516-034226

Pion, G. M., & Lipsey, M. W. (2021). Impact of the Tennessee Voluntary Prekindergarten Program on children's literacy, language, and mathematics skills: Results from a regression-discontinuity design. *AERA Open, 7*. https://doi.org/10.1177/23328584211041353

Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J. R., Porter, J., & Smith, J. (2010). *Standards for regression discontinuity designs*. What Works Clearinghouse. https://ies.ed.gov/ncee/wwc/Document/231

Sekhon, J. S. (2011). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software 42*(7), 1–52. https://doi.org/10.18637/jss.v042.i07

Shadish, W. R. (2002). Revisiting field experimentation: Field notes for the future. *Psychological Methods, 7*(1), 3–18. https://doi.org/10.1037/1082-989X.7.1.3

Smith, A., Staunton, R., Sahasranaman, A., & Worth, J. (2023). *Impact evaluation of Nuffield Early Language Intervention (NELI) Wave Two: Evaluation report*. Education Endowment Foundation. https://www.nfer.ac.uk/media/g0wlav3a/impact_evaluation_of_nuffield_early_language_intervention_wave_two_evaluation_report.pdf [PDF]

Snowling, M. J., West, G., Fricke, S., Bowyer-Crane, C., Dilnot, J., Cripps, D., Nash, M., & Hulme, C. (2022). Delivering language intervention at scale: Promises and pitfalls. *Journal of Research in Reading, 45*(3), 342–366. https://doi.org/10.1111/1467-9817.12391

Strand, S., & Hessel, A. (2018). *English as an additional language, proficiency in English and pupils' educational achievement: An analysis of local authority data.* Bell Foundation.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*(1), 1–21. https://doi.org/10.1214/09-STS313

Trochim, W. M., & Donnelly, J. P. (2001). *Research methods knowledge base* (Vol. 2). Atomic Dog Publishing.

Waddington, H. S., Villar, P. F., & Valentine, J. C. (2023). Can non-randomised studies of interventions provide unbiased effect estimates? A systematic review of internal replication studies. *Evaluation Review, 47*(3), 563–593. https://doi.org/10.1177/0193841X221116721

West, G., Snowling, M. J., Lervåg, A., Buchanan-Worster, E., Duta, M., Hall, A., McLachlan, H., & Hulme, C. (2021). Early language screening and intervention can be delivered successfully at scale: Evidence from a cluster randomized controlled trial. *Journal of Child Psychology and Psychiatry, 62*(12), 1425–1434. https://doi.org/10.1111/jcpp.13415