Hanen Learning Language and Loving It™

Statistical Analysis Plan

NatCen Social Research

Principal investigator(s): Jonah Bury

Education Endowment Foundation

Template last updated: August 2019

PROJECT TITLE[1]	Hanen Learning Language and Loving It (Hanen LLLI)
DEVELOPER (INSTITUTION)	The Hanen Centre
EVALUATOR (INSTITUTION)	National Centre for Social Research
PRINCIPAL INVESTIGATOR(S)	Jonah Bury
PROTOCOL AUTHOR(S)	Sashka Dimova, Molly Scott
TRIAL DESIGN	Two-arm cluster randomised controlled trial with random allocation at Early Years setting level
TRIAL TYPE	Efficacy
PUPIL AGE RANGE AND KEY STAGE	3 to 4 years old, Early years
NUMBER OF SETTINGS	140
NUMBER OF PUPILS	2,099
PRIMARY OUTCOME MEASURE AND SOURCE	Receptive language measured with the British Picture Vocabulary Scale (BPVS), Third Edition, GL Assessment.
SECONDARY OUTCOME MEASURE AND SOURCE	1. Expressive and receptive language measured with the Renfrew Action Picture Test (RAPT), 5th Edition, Routledge 2. Socio-emotional behaviour measured with the Strengths and Difficulties Questionnaire – Teacher version (SDQ-T) (Goodman, 2001)

SAP version history

VERSION	DATE	REASON FOR REVISION
1.1		

[1] Make sure that the project title here matches the title of the document and the protocol. Please ensure that there is an identification as a randomised trial in the title as per CONSORT requirements.

1.0 [*original*]	25 November, 2022	*N/A*

Table of Contents

Abbreviations

BPVS	British Picture Vocabulary Scale
CACE	Complier Average Causal Effect
CPD	Continuing Professional Development
DfE	Department for Education
EAL	English as an Additional Language
EEF	Education Endowment Foundation
EY	Early Years
EYPP	Early Years Pupil Premium
FSM	Free School Meals
Hanen LLLI	Hanen Learning Language and Loving It
ICC	Intra-Cluster Correlation
ITT	Intention to Treat
IPE	Implementation and Process Evaluation
MCAR	Missing Completely at Random
MDES	Minimum Detectable Effect Size
MI	Multiple Imputation
MICE	Multiple Imputation using Chained Equations
OLS	Ordinary Least Squares
PVI	Private, Voluntary and Independent Sector
RAPT	Renfrew Action Picture Test
RCT	Randomised Controlled Trial
SAP	Statistical Analysis Plan
SDQ-T	Strengths and Difficulties Questionnaire (Teacher version)
SEN	Special Educational Needs
SLT	Speech and Language Therapist
ToC	Theory of Change
URN	Unique Reference Number

Introduction

This Statistical Analysis Plan (SAP) describes the planned impact evaluation analysis for an efficacy trial of Learning Language and Loving It™ – The Hanen Program® for Early Childhood Educators (Hanen LLLI).

Hanen LLLI is a Continuing Professional Development (CPD) program for Early Years (EY) practitioners designed to provide staff with practical strategies to enhance children's communication and language skills through specialised ways of interacting and communicating with children during normal daily routines.

It was developed by the Hanen Centre[2], in Canada and previous evidence from randomised controlled trials (RCTs) from the US suggests that Hanen LLLI can have a positive impact on children's language attainment (see Piasta et al., 2012; Law et al., 2017).

A small-scale pilot evaluation of Hanen LLLI conducted by NatCen Social Research in 2019 found the intervention to be attractive to settings and showed evidence of promise regarding changes to nursery staff's interactions with children. However, no impact evaluations of Hanen LLLI have taken place in the UK to date.

The evidence base relating to the potential impact of Hanen LLLI prompted the Education Endowment Foundation (EEF) to commission this evaluation. This evaluation was commissioned as part of the Department for Education's (DfE) Accelerator Fund. The EEF previously commissioned an efficacy trial of Hanen LLLI; however, the impact evaluation was cancelled as post-trial data collection was not feasible during the COVID-19 pandemic.

This evaluation aims to provide evidence of the impact of Hanen LLLI in the UK context.

Intervention overview

To enable effective delivery, Hanen LLLI includes a number of activities:

- An initial information event for settings;
- One introductory workshop for practitioners to explain the intervention and evaluation[3];
- Eight training workshops for practitioners lasting 2.5 hours each;
- Six individual video feedback sessions lasting 30 to 40 minutes;
- A pre-intervention (baseline) and a post-intervention (endline) video of setting staff recording their practice with children;
- One baseline visit and one post-programme visit per setting, carried out in person.

Through participation in the programme activities, practitioners learn practical, research-based strategies for engaging with children to enhance their language development and social skills.

[2] The Hanen Centre's mission is to enable parents and professionals to transform their daily interactions with young children to build the best possible lifelong social, language and literacy skills.
[3] The Hanen Centre uses the term 'orientation meetings' instead of 'introductory workshops'.

The Hanen Centre recommend that workshops are delivered to groups of 10-20 practitioners and that video feedback sessions are delivered one-to-one. The training workshops and video feedback sessions are delivered by qualified and certified Hanen LLLI speech and language therapists (SLTs) and EY Teachers also known as Program Leaders. Program Leaders are expected to facilitate the four broad aims of Hanen LLLI: education; application; collaboration and peer support.

After the disruption caused by the COVID-19 pandemic, delivery of the training workshops was adapted to include a mixture of in-person and online modes. In this trial, training workshops 1, 2, 5, 6, and 8 are completed in-person, while the remaining training workshops and all video feedback sessions take place online.

The formal and informal cascading of learning is a key aspect of the Hanen LLLI programme being implemented for this trial. The formal cascading of training content is carried out using responsive interaction strategies[4] and provision of information, which enhances buy-in from senior leadership and participants. Informal cascading is enabled through the mixing of trained and non-trained practitioners within the setting and encouraging information sharing between these two groups.

During the baseline visit, Program Leaders visit each setting to ensure they have a sufficient understanding of the programme and to encourage engagement and cascading. During the post-programme visit, Program Leaders return to each setting to reflect on their learning and discuss future plans for using Hanen strategies and cascading learning.

For more information regarding Hanen LLLI, its background and design, please refer to the evaluation protocol (Dimova et al., 2022).

Evaluation overview

In this trial, the programme will be delivered over the course of 31 weeks between November and June in the 2022/23 school year. In total, 167 settings were initially recruited and randomised. 140 settings remain in the trial after baseline assessment, of which 72 were assigned to receive Hanen LLLI and 68 were assigned to a control group (teaching as usual)[5]. The trial includes school-based maintained settings and settings from the Private, Voluntary and Independent (PVI) sector. The inclusion of PVIs in this evaluation was motivated by the fact that many children from disadvantaged backgrounds receive EY education in PVI settings, but the available evidence base for this sector has typically been weaker in comparison to maintained nurseries.

The trial is taking place in three Regional School Commissioner areas: The North (covering Cumbria, the North East and North Yorkshire), East Midlands and the Humber, and the West Midlands.

This evaluation is based on a logic model, which sets out the intended impacts of Hanen LLLI for settings and children, and the short, medium and long-term outcomes for staff, settings and children that are expected to lead to these impacts.[6] The intervention logic

[4] More information on the responsive interaction strategies can be found on the Hanen webpage.
[5] Note that 5 of the 72 settings assigned to receive Hanen LLLI are not ultimately participating in the programme, and will be considered as 'non-compliers'. This is discussed further below.
[6] The logic model does not include contextual variables and causal mechanisms. These will be explored as part of IPE activities.

model had been initially developed for Hanen LLLI 1[7] in conjunction with Communicate SLT, and this has been further updated for this trial during the Theory of Change (ToC) workshop (see Bury et. al, 2022). The focus of the impact evaluation in this trial is to assess if Hanen LLLI leads to improvements in i) children's receptive and expressive language and ii) socio-emotional behaviour. The target group are children 3-4 years old in settings that agreed to take part in this trial.

For this trial, Hanen LLLI is delivered by Communicate SLT CIC[8], and is independently evaluated by NatCen. The study is funded by the EEF.

Design overview

The impact evaluation is designed as a stratified, two-arm cluster randomised efficacy trial, with settings as the unit of randomisation and pupils as the unit of analysis.

167 eligible settings were recruited for the trial and were randomised to receive Hanen LLLI or not on the 25th August 2022. We randomised settings within regions and by setting type (maintained vs PVI settings). Settings had a 50:50 chance of being assigned to the treatment group within each region-setting strata. Settings were recommended to sign up two thirds of staff working with 3- to 4-year-old children to take part in the programme activities. Within recruited settings, up to 17 eligible children were randomly selected to take part in baseline and endline outcome testing. The decision to test up to 17 children was informed by budget considerations and power calculations (see Sample size calculations overview). If settings had less than 17 children, then all children eligible for Hanen LLLI were assessed at baseline. We also randomly selected an additional three pupils (or up to three, depending on the size of the setting), to act as replacements for any sampled pupils who were absent during the baseline testing. Children were randomly sampled for assessment by a member of the evaluation team. The same children who were selected for baseline assessment will complete the endline assessment. The full list of exclusion criteria for settings and pupils in the trial are outlined in the protocol.

There is only one treatment condition: settings allocated to the treatment group are offered the chance to take part in Hanen LLLI activities throughout the 2022/23 academic year. Settings assigned to the Hanen LLLI group will receive the programme for free, but they will not receive any monetary incentives[9]. Settings assigned to the control condition will implement their 'business as usual' approach to language teaching. As an incentive for participation in data collection activities, control settings will receive two payments: a payment of £100 for completion of baseline testing in November 2022 and £900 in August 2023 for completion of endline testing and for completing and sending videos to monitor changes in staff's practice. The incentive is intended to mitigate the risk that settings engage when approached about the trial but choose not to participate once assigned to the control group.

[7] For more information on Hanen LLLI see https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/learning-language-and-loving-it-efficacy
[8] A Community Interest Company who provide Speech and Language Therapy services, based in the North West of England. Communicate SLT are Hanen-certified trainers for some of the Hanen programs but are otherwise not affiliated in any way with The Hanen Centre.
[9] Settings received a payment of 75% of their cover cost to provide for staff cover while practitioners attended the training.

The outcome measures selected for the evaluation are age-appropriate, fit well with the Hanen LLLI logic model and were selected in collaboration with the delivery team. The primary outcome of interest is receptive language as measured by the age standardised British Picture Vocabulary Scale, Third Edition (BPVS-3). The first secondary outcome measure provides a further measure of receptive and expressive language, as measured by the Renfrew Action Picture Test (RAPT). In addition to language, the evaluation will also assess differences in socio-emotional behaviour using the Teacher version of the Strengths and Difficulties Questionnaire (SDQ-T).

Settings in the trial completed the baseline assessment in October 2022, and will undertake the endline assessment between June and July 2023. Baseline assessment took place following randomisation, but before assignment to treatment or control group was communicated to settings. In total, 140 settings out of the 167 that were randomised completed the baseline assessment (72 treatment and 68 control). 27 settings withdrew between randomisation and baseline assessment for various reasons. The reasons for withdrawal were: staffing issues (N=14), timescales of the evaluation activities (N=6); or settings being unable to upload pupil data in time (N=6). There was one further setting that withdrew during the baseline assessment, after some pupils had already been assessed. This setting has been removed from the trial, and the assessment data that was already been collected for some pupils has been deleted.

At the time of writing (March 2023) Hanen LLLI was being delivered in 67 settings. 5 settings that were initially randomised to take part in Hanen LLLI are not ultimately receiving the programme. These settings will be considered 'non-compliers'. This means that they will still be included as part of the Hanen treatment group for the purpose of impact estimation, even though they have not taken part in practice.

Table 1 summarises the trial design.

Table 1 Trial design

Trial design, including number of arms		Two-arm cluster randomised trial
Unit of randomisation		Setting
Stratification variables (if applicable)		Geographic region
Primary outcome	variable	Receptive language
	measure (instrument, scale, source)	Receptive language measured with the British Picture Vocabulary Scale Third Edition (BPVS-3), age standardised score (85-115)[10], GL Assessment
Secondary outcome(s)	variables	1. Expressive and receptive language 2. Socio-emotional behaviour

[10] The age-expected range of standardised BPVS scores is 85-115. The full possible range of scores is 69-141.

	measures (instrument, scale, source)	1. Renfrew Action Picture Test (RAPT), 5th Edition, Routledge (RAPT) Information and Grammar, raw score (0-41 Information; and 0-39 Grammar), Routledge 2. The Strengths and Difficulties Questionnaire -Teacher version (SDQ-T), 0-40, Goodman (1998, 2001)
Baseline for primary outcome	**variable** measure (instrument, scale, source)	Receptive language BPVS-3 age-standardised score, 85-115, GL Assessment
Baseline for secondary outcome	**variable** measure (instrument, scale, source)	1. Expressive and receptive language 2. Socio-emotional behaviour 1. BPVS-3 age standardised score, 85-115, GL Assessment 2. Not applicable (post-test only)

Research questions

This impact evaluation aims to answer the following research questions (RQ):

RQ1: To what extent did Hanen LLLI lead to changes in children's receptive language as measured by the BPVS? (Primary outcome)

RQ2: To what extent did Hanen LLLI lead to changes in children's receptive and expressive language as measured by the RAPT? (Secondary outcome)

RQ3: To what extent did Hanen LLLI lead to changes in children's socio-emotional development as measured by the SDQ-T? (Secondary outcome)

RQ4: To what extent did Hanen LLLI lead to changes in receptive language as measured by the BPVS for children who are entitled to Early Years Pupil Premium (EYPP)? (Subgroup analysis)

RQ5: To what extent did Hanen LLLI lead to changes in receptive language for lower and higher ability pupils based on the BPVS? (Subgroup analysis)

Randomisation

Randomisation of settings was carried out by a member of the evaluation team on 25 August 2022, in Stata version 17. The researcher carrying out randomisation was blinded to setting identity at the time of carrying out the randomisation. This means that meaningful identifiers such as settings names and Unique Reference Numbers (URNs) were removed at the time of randomisation, and later linked back.

The randomisation was stratified by region and setting type (maintained vs PVI), to help ensure that settings from the same region, or with a similar type of provision, were evenly allocated to the treatment and control groups. We used the *randtreat* command in Stata to perform randomisation, using the misfits(global) option.

Altogether, 167 settings were randomised. 83 were assigned to the treatment group and 84 to the control group. Table 2 shows the treatment allocation within each stratum. The number of settings varied across strata, from 9 PVI settings in the North of England, to 56 school-maintained settings in the West Midlands. The randomisation produced an equal allocation to the treatment and control group within each stratum, with the exception for PVIs in the North of England where there was an uneven number of settings within the stratum.

Table 2 Randomisation allocation across strata (Area, Setting type)

Strata	Treatment	Control
East Midlands and the Humber, Maintained	13	13
East Midlands and the Humber, PVIs	14	14
North of England, Maintained	14	14
North of England, PVIs	4	5
West Midlands, Maintained	28	28
West Midlands, PVISs	10	10

The allocation by region and programme area (see Table *3* and Table 4) was recorded in Excel and communicated to the delivery team on 26 August 2022 to facilitate planning for the training.

As specified in the study protocol, settings were informed of their allocation status after baseline data collection was completed in October 2022.

Table 3 Number of settings allocated to Hanen LLLI by area

Area	Treatment	Control
West Midlands	38	38
East Midlands and the Humber	27	27
North of England	18	19

Table 4 Number of settings allocated to Hanen LLLI by programme area

Programme area	Treatment	Control
Birmingham	24	25
Cumbria	8	8
Derby	6	11
Leicester	12	10
Newcastle	9	11
Sheffield	10	6
Telford	5	6
Worcestershire	9	7

Primary outcome

The primary outcome for the trial is English receptive language, measured using the BPVS-3. BPVS-3 is an individually administered, norm-referenced, test of receptive vocabulary for Standard English. We will use the standardised BPVS-3 score for the primary outcome, constructed using conversion tables available in the BPVS-3 scoring manual. The standardised scores reflect how a child's score compares to other children of a similar age in months, based on a large nationally representative standardisation sample. The score is standardised to a mean of 100 indicating whether a child is above or below the BPVS's national average score. Conversion tables used to standardise the raw BPVS scores are available in the BPVS-3 Manual.

The BPVS-3 has several advantages. One advantage is that it is quick to administer and score. It also has been standardised with a UK sample and has relatively high validity and reliability[11]. However, the test is only suitable for measuring hearing vocabulary at a particular point in time. For this reason, it is advised to avoid over-generalising conclusions on learning outcomes (Dunn et. al. 2009).

BASELINE ASSESSMENT

Parents and caregivers were provided with information about the trial before baseline assessments took place, and they had the option to opt out on behalf of their child. Of children who were eligible for Hanen LLLI who were not opted out, we sampled up to 17 children per setting for baseline assessment on the 22nd September 2022. If settings had fewer than 17 children who had not been opted out of the trial, then all children were included in the data collection. In larger settings with more than 17 children, we also sampled up to three additional children (depending on the size of the setting) to serve as replacements for any selected children that will be absent on the day or unable to undertake the baseline assessment for any other reason.

Baseline BPVS assessments were carried out in October 2022 by SLTs trained in administering the BPVS. The same children will be invited to complete endline BPVS assessments towards the end of the summer term 2023.

Secondary outcomes

The trial includes two secondary outcomes that will be measured at endline only.

RECEPTIVE AND EXPRESSIVE LANGUAGE

The first secondary outcome is the Renfrew Action Picture Test (RAPT), which measures receptive and expressive language. The RAPT comprises 10 pictures depicting various scenarios. Children's receptive and expressive language is tested by asking children to describe the pictures that they are shown. Children's answers are recorded and then scored according to two separate perspectives: information and grammar. The main advantage of the RAPT is that it provides a snapshot of a child's level of expressive language. However, as with the BPVS, the results represent a snapshot that should be interpreted as part of a wider assessment of language abilities.

We will use the RAPT raw score because the test has been standardised only on children who speak English as a first language, so the standardised score will be misleading if applied to children who speak English as an additional language (EAL). The raw score for

[11] For more information see: https://educationendowmentfoundation.org.uk/measures-database/british-picture-vocabulary-scale

information ranges from 0 to 41, while the raw score for grammar ranges from 0 to 39 (Renfrew, 1988).

The RAPT will be administered by SLTs trained to administer the test who will not be informed of settings' treatment allocation at the time of assessment.

SOCIO-EMOTIONAL BEHAVIOUR

The trial also includes an additional secondary outcome; - socio-emotional behaviour, as measured by the teacher version of the SDQ. This will be assessed at endline by setting staff who know the children well. The SDQ is a brief emotional and behavioural screening measure. The measure is well established and has good validity and reliability (Goodman et al, 1998; Goodman, 2001). The SDQ comprises 25 items. The questionnaire is divided into five subscales measuring emotional symptoms, conduct problems, hyperactivity/inattention, peer relationship problems and prosocial behaviour. It provides a 'total difficulties score', by summing scores from 4 out of the 5 sub-scales (excluding the prosocial subscale). The total difficulties score ranges from 0 to 40. This trial will use the SDQ raw scores.

Timeline

Table 5 summarises key dates for the impact evaluation. A full timeline for the efficacy trial, covering all evaluation activities, can be found in the evaluation protocol.

Table 5 Timeline of key dates for the impact evaluation

Date	Activity
February – July 2022	Recruitment of settings
25 August 2022	Randomisation of settings
September 2022	Children data provided by settings for all 3–4-year-olds who were not opted out.
22 September	Random selection of 17 eligible children for baseline and endline assessments
3 – 21 October 2022	Baseline BPVS assessments carried out with sampled pupils
End of October 2022	Settings informed of randomisation allocation (after testing)
November 2022	Settings in control group receive payment of £100 for completion of baseline testing
November 2022 – June 2023	Hanen LLLI implementation delivery in the treatment group
July 2023	Endline pupil assessments (BPVS, RAPT and SDQ-T)
August 2023	Control group settings receive second payment of £900 for completion of endline assessments
October 2023	Submission of draft EEF report
January 2024	Final EEF report

Sample size calculations overview

We have used PowerUp! (Dong and Maynard, 2013) to perform all sample size calculations.

Planned sample sizes

The evaluation protocol anticipated the following sample sizes:

- We aimed to recruit 165 settings in the trial, with half randomly allocated to the treatment group and half to the control group. We further agreed with Communicate SLT that roughly 60-70% of recruited settings should be maintained nurseries, and 30-40% should be PVIs.
- Within each setting we aimed to carry out baseline BPVS assessments with up to 17 pupils, depending on the size of the setting.

This sampling strategy was designed with the expectation that roughly 20% of the recruited settings would be likely to withdraw from the trial after initial recruitment. We also anticipated that some settings would have fewer than 17 pupils in total (since settings could be eligible for the trial if they had a minimum of 12 pupils aged 3-4).

In view of this expected attrition, and the inclusion of smaller settings, we anticipated that the sampling strategy outlined in the protocol would be likely to result in a sample size of about 132 settings by the time of endline testing, with an average of just less than 15 pupils per setting. This implies a planned sample size of around 1,980 pupils altogether by the time of endline testing.

The estimated Minimum Detectable Effect Size (MDES) associated with our planned sample size was 0.217 standard deviations. The MDES captures the smallest possible change in outcomes that the trial is likely to be able to detect. 0.217 is slightly less favourable than the MDES required for the trial to receive a '5-padlock' rating by the EEF.

Our sample size calculations and the assumptions that underpin them are described in the next sub-section and summarised in Table 7 below.

Achieved sample sizes so far

This SAP was written shortly after the completion of baseline testing.

Below we summarise the sample sizes that have been achieved at each stage of the trial so far.

- **By the time of randomisation (August 2022):** 167 settings were successfully recruited into the trial by August 2022 and randomised into either the Hanen LLLI treatment or business-as-usual control group. 66% of these were maintained settings and 34% were PVIs.
- **By the time of pupil enumeration (September 2022):** 143 settings remained in the trial and successfully enumerated all their nursery-age pupils in preparation for baseline data collection. Settings that dropped out between randomisation and pupil enumeration are not considered to be part of the trial, i.e., they are not eligible to take part in Hanen LLLI or to receive any incentive.
- **Baseline assessments (October 2022):** 140 settings successfully completed baseline BPVS assessments, with a total of 2,099 pupils assessed (an average of around 15 pupils assessed per setting). Two settings allocated to the treatment group did not complete the baseline assessment due to staffing issues, and one further setting withdrew part of the way through the testing period. These settings are no longer considered part of the trial.

At baseline, the final data revealed a deviation from the planned sampling approach. Although we planned to carry out assessments with up to 17 pupils per setting (with up to 3 replacement pupils also identified in case of absence), in 15 settings we found that more than 17 pupils were assessed. In these settings, the replacement pupils were assessed as well as the intended sampled pupils. We have retained the full sample of pupils assessed in the trial, including those who were not intended to be assessed, to avoid discarding data already collected. We will aim to carry out endline assessments with all pupils assessed at baseline.

The sample breakdown by treatment group, at each stage of the trial so far, is summarised in Table 6. We anticipate that there is likely to be some further attrition by the time of the endline assessments in 2023.

Table 6 Achieved sample sizes at each stage of the trial so far

Trial stage	Number of settings		Number of pupils	
	Treatment	Control	Treatment	Control
By the time of randomisation (August 2022)	83	84	N/A	N/A
By the time of pupil enumeration (September 2022)	74	69	1,206	1,099
Baseline assessments (October 2022)	72	68	1,104	995

Updated sample size calculations

Table 7 presents updated sample size calculations for this trial. These calculations indicate the smallest effect size, measured in standard deviations, that the trial is likely to be able to detect given its sample sizes and a set of underlying assumptions (i.e., it's MDES).

The revised calculations are based on the number of settings and pupils that have completed baseline data collection; these represent the sample sizes that remain in the trial at the time of writing. The calculations use the same core assumptions as those conducted in the evaluation protocol. As before, we assume a high pupil-level correlation between baseline and follow-up (0.60) and moderate setting-level correlation (0.36)[12]. We use a Type I error rate of 0.05 and a Type II error rate of 0.20 (i.e., power of 0.80).

As well as updating the expected number of pupils and settings in the trial, we have also updated some of the other assumptions underpinning these calculations. We use an Intra-Cluster correlation of 0.171, based on what we found in the baseline assessment data. This is slightly more favourable than the ICC of 0.185 used in the protocol calculations (which were themselves based on the 2019 baseline data collected). In this case, the ICC of 0.171 is somewhat lower than what was previously assumed, leading to gains to the expected MDES.

Note that this assumption on the ICC is still conservative as it is unadjusted. This means that it simply captures the degree of association in baseline outcomes between pupils belonging to the same setting. The ICC that will matter in practice for the precision of final impact estimation will in fact be the residual ICC that remains after adjusting for baseline attainment and strata (which are the covariates that will be included in our final analysis model). If

[12] Power calculations use variance explained by covariates (R^2) as opposed to pre-post-test correlations. We approximate the pre-post-correlation by taking the square root of the R^2.

baseline attainment and strata help to explain some of the variance in final outcomes, then the ICC that remains after adjusting for these covariates will be lower than the unadjusted ICC. For that reason, the simple ICC calculated in the baseline data is a conservative one.

While we have updated total expected sample size of pupils and settings, based on the baseline data, we have not updated our assumptions about the number of EYPP pupils we might expect to find within this sample for our subgroup analysis. This is because information about pupils' EYPP status is still being gathered by participating settings at the time of writing, so we do not have complete information about this yet. We therefore continue to use an assumption of around 2 pupils per setting who will be eligible for EYPP.

Using these updated assumptions, we find that the current sample sizes would yield an MDES of 0.206 for the primary analysis, and an MDES of 0.307 for the sub-group analysis by EYPP status. The MDES of 0.203 is somewhat lower than our initial expectation and is closer to the threshold required for the trial to receive a '5-padlocks' security rating. This indicates that the trial is so far on track to at least meet the MDES assumptions outlined in the protocol, even if there is further attrition by the time of the endline data collection.

Table 7 Sample size calculations

		Protocol		SAP[1]	
		OVERALL	EYPP	OVERALL	EYPP
Minimum Detectable Effect Size (MDES)		0.217	0.322	0.206	0.307
Pre-test/ post-test correlations	level 1 (pupil)	0.60	0.60	0.60	0.60
	level 2 (setting)	0.36	0.36	0.36	0.36
Intra-cluster correlation (ICC)	level 2 (setting)	0.185	0.185	0.171	0.171
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		2	2	2	2
Average cluster size		15	2	14[2]	2
Number of settings	treatment	66	66	72	72
	control	66	66	68	68
	Total	132	132	140	140
Number of pupils	treatment	990	132	1104	144
	control	990	132	995	138
	Total	1980	264	2099	282

Notes: Power calculations were performed using PowerUp! under an alpha level of 0.05 and power of 0.8. The calculations include estimates of the proportion of variance explained through the included covariates at each of these levels (also known as R^2). The R^2 values here have been estimated by squaring the pre-test post-test correlation. An R^2 value of 0.36 at pupil-level and 0.13 at setting-level is used in the power calculations. Here we use EYPP here rather than Free School Meal (FSM) status, as EYPP data is directly available from settings, whereas FSM is not. *Power calculations at the SAP stage are based on the actual achieved sample following baseline assessment. **The average cluster size of 14 is the harmonic mean of the number of pupils assessed per setting. The arithmetic mean is 15. Harmonic means are recommended for use in sample size calculations where cluster sizes vary, since it is less sensitive to outliers and is therefore more conservative than other types of means.

Analysis

Primary outcome analysis

RQ1: To what extent did Hanen LLLI lead to changes in children's receptive language as measured by the BPVS?

The primary outcome analysis will use an Intention-to-Treat (ITT) approach to estimate the impact of Hanen LLLI on children's language development. This addresses research question 1. The model will be a two-level linear regression, with pupils at level one and settings at level two. We are using a multilevel model to account for the fact that the trial is designed as a cluster randomised controlled trial, with pupils clustered within settings.

The BPVS standardised score at follow-up will be the dependent variable, with a binary indicator of treatment allocation, baseline BPVS score, and geographic region-setting type (the randomisation strata) included as independent variables. Setting-level random effects will be included in the model by allowing the intercept to vary by setting. This implies the following model:

$$(1)\ BPVS_{ij} = \beta_0 + \beta_1 Intervention_j + \beta_2 Stratum + \beta_3 Baseline\_BPVS_{ij} + u_j + e_{ij}$$

Where pupils (i) are nested within settings (j). The intervention effect is estimated by β_1, whilst u_j represents the setting random effect and e_{ij} represents the individual error term. In line with EEF guidance (EEF, 2018a) no further covariates will be included.

Equation (1) is known as a 'random intercepts' model because $\beta_{0j} = \beta_0 + u_j$ is interpreted as the setting specific intercept for setting j and $\beta_{0j} \sim i.i.d\ N(\beta_0, \sigma_u^2)$ is random, i.e., it is a number that can take any value.

The target parameter that will indicate the average effect of the intervention on children in treatment settings compared to those in control settings is β_1.

The analysis will be conducted in Stata 17 using the mixed command.

Secondary outcome analysis

RQ2: To what extent did Hanen LLLI lead to changes in children's receptive and expressive language as measured by the RAPT?

RQ3: To what extent did Hanen LLLI lead to changes in children's socio-emotional development as measured by the SDQ-T?

We will estimate the impact of Hanen LLLI on two secondary outcomes: the RAPT and SDQ-T (see Secondary outcomes).

The analytical approach for secondary outcome analysis will be analogous to the primary outcome estimation. We will use an ITT approach, consisting of a multi-level linear regression model with pupils nested within settings. Each secondary outcome as measured at endline will be regressed on a binary indicator of treatment allocation and the stratum to which the setting belongs. Setting-level random effects will be included to account for the variance at setting level.

In the case of the RAPT analysis, we will include baseline BPVS scores as an additional covariate. This is because we do not have a baseline measure for the RAPT; however, baseline BPVS scores are expected to have predictive power for the RAPT at endline. Using a different baseline measure of attainment may lower the pre- post- test correlation, reducing

the power of the secondary analysis. We will omit this covariate from the SDQ-T outcome model.

This implies the following models for secondary analysis:

(2) $RAPT_{ij} = \beta_0 + \beta_1 Intervention_j + \beta_2 Stratum + \beta_3 Baseline\_BPVS_{ij} + u_j + e_{ij}$

(3) $SDQ_{ij} == \beta_0 + \beta_1 Intervention_j + \beta_2 Stratum + u_j + e_{ij}$

Subgroup analyses

RQ4: To what extent did Hanen LLLI lead to changes in receptive language as measured by the BPVS for children who are entitled to EYPP?

RQ5: To what extent did Hanen LLLI lead to changes in receptive language for lower and higher ability pupils based on the BPVS?

We will conduct two sub-group analyses for this trial as defined in the study protocol. The subgroup analysis are exploratory as the study is not powered for meaningful sub-group analysis[13].

The first will assess whether Hanen LLLI has a differential impact for pupils who are eligible for the EYPP in the treatment group compared to pupils eligible for EYPP in the control group.[14] This addresses research question 4. The EYPP provides settings with additional funding for all 3–4-year-olds from low-income families. In order to carry out this analysis, we asked settings to provide information about the EYPP status of all participating children as part of the pupil enumeration exercise carried out in September 2022.

The second sub-group analysis will examine differential impacts for children with lower initial language development who were in Hanen LLLI settings and those with lower language development in control settings. This addresses research question 5, which aims to explore whether Hanen LLLI helps pupils starting from a lower base to 'catch up' with their peers. We will construct the sub-groups for this analysis by identifying pupils scoring in the bottom quartile of the age standardised BPVS assessment at baseline. We will construct a binary indicator to define low language achievers as those who scored below the threshold.

We will carry out both sub-group analyses using a similar approach to the primary analysis. We will fit a multilevel linear regression model, with the BPVS standardised score as the independent variable. This will be regressed on a binary indicator of treatment allocation, the sub-group indicator, an interaction term for treatment allocation and the sub-group indicator, stratum, baseline BPVS attainment and a random effect for settings.

(4) $BPVS_{ij} = \beta_0 + \beta_1 Intervention_j + \beta_2 EYPP_{ij} + \beta_3 Intervention_j * EYPP_{ij} + \beta_4 Stratum + \beta_5 Baseline\_BPVS_{ij} + u_j + e_{ij}$

[13] We estimated an MDES of 0.307 for the sub-group analysis according to pupil Early Years Pupil Premium (EYPP) eligibility status.

[14] We note that EEF protocols usually include sub-group analysis by Free School Meal (FSM) status. However, we use EYPP here in preference to FSM as EYPP data is available directly from settings, whereas FSM is not. EYPP provides schools with additional funding for all 3-4 year-olds from low-income families. All EYPP pupils are also eligible for FSM.

(5) $BPVS_{ij} = \beta_0 + \beta_1 Intervention_j + \beta_2 Low\_ability_{ij} + \beta_3 Intervention_j * Low\_ability_{ij} + \beta_4 Stratum + \beta_5 Baseline\_BPVS_{ij} + u_j + e_{ij}$

In each model, the interaction term (β_3) describes the differential impact of Hanen LLLI for each sub-group. So, in equation 4, β_3 indicates whether there is a differential impact of Hanen for EYPP pupils, and in equation 5, β_3 will show whether there is a differential impact of Hanen LLLI for pupils with lower initial language development.

If the interaction term coefficient is statistically significant, an additional model will be estimated. This model will take the same form as the primary analysis model (equation 1), using solely the sample of pupils who are part of each sub-group. That is the model will be re-estimated for only the sample of pupils eligible for EYPP, or only the sample who were designed as having lower initial language development at baseline.

Additional analyses

We will not carry out any additional analyses.

Longitudinal follow-up analyses

Long term follow-up for this trial is currently not considered and thus not discussed further in this SAP.

Imbalance at baseline

To check for, and monitor, imbalance between treatment and control groups at baseline following baseline assessment, descriptive analysis was undertaken at the setting and child level.

At the setting level, we looked at the following variables, by means of cross-tabulation within the control and treatment groups from the data in the study sample (rather than publicly available statistics):

- Proportion of PVIs
- Proportion across RSC area
- Number of pupils enumerated

At the individual level, balance was assessed for the following characteristics:

- EYPP status
- Child's age in months
- Standardised BPVS score

Later in the study we will assess balance on the following characteristics:

At the setting level:

- Ofsted rating
- Proportion of EYPP children in the setting
- Average scaled score in KS2 reading for school-based settings

At the time of writing this SAP (November 2022), pupil enumeration information and baseline assessment data has already been collected in 140 participating settings. We are therefore able to present findings on imbalance at baseline at this stage. These are shown in Table 8

below. Overall, the treatment and control groups appear very similar at baseline. This is expected given the random assignment.

33% of treatment settings and 35% of control settings are PVIs. This is in line with the recruitment target outlined in the evaluation protocol. Settings are fairly evenly distributed across the three RSC areas, with the highest proportion of settings in the West Midlands, and the lowest in the North of England. Settings enumerated an average of around 29 pupils each (not all of whom were sampled for baseline assessment).

In total, 16% of children in the treatment group are EYPP pupils, compared to 19% in the control group. At this stage there remain some missing values for EYPP status. During the time of the pupil enumeration, some settings reported that they do not record EYPP status until later in the year. A mop-up exercise was carried out in October and November to recover missing EYPP status (and some other missing children information). However there remain some children whose EYPP status is unknown at this stage. This will be collated before endline assessment.

Children were an average of 43 months old at the time of the baseline assessment across treatment and control settings.

Children in both groups had an average BPVS score of 91. The BPVS instrument has been standardised such that the national average score for each age group is 100. Standardised scores of below 100 indicate that children are attaining a lower level of receptive language development, on average, than the national average for their age. This is consistent with the aims of Hanen LLLI to support settings that have a relatively higher need for the intervention.

Table 8 Baseline characteristics of groups as randomised

Setting-level (categorical)	Treatment group		Control group	
	n/N (missing)	Count (%)	n/N (missing)	Count (%)
PVIs	24 /72 (0 missing)	33.33	24 /68 (0 missing)	35.29
RSC area: East Midlands & the Humber	23 /72 (0 missing)	31.94	22 /68 (0 missing)	32.35
RSC area: North of England	17 /72 (0 missing)	23.61	17 /68 (0 missing)	25
RSC area: West Midlands	32 /72 (0 missing)	44.44	29 /68 (0 missing)	42.65
Setting-level (continuous)	**n/N (missing)**	**Mean (SD)**	**n/N (missing)**	**Mean (SD)**
Number of pupils enumerated	Total observations = 72/72 (0 missing)	29.01 (15.06)	Total observations = 68/68 (0 missing)	27.54 (14.45)
Child-level (categorical)	**n/N (missing)**	**Count (%)**	**n/N (missing)**	**Count (%)**
EYPP status	161 /1104 (106 missing)	16.13	174 /995 (67 missing)	18.75
Child-level (continuous)	**n/N (missing)**	**Mean (SD)**	**n/N (missing)**	**Mean (SD)**
Child age in months	Total observations = 1104/1104	43.01 (3.93)	Total observations = 994/995	43.28 (3.96)

	(0 missing)		(1 missing)	
Standardised BPVS score	Total observations = 1089/1104 (15 missing)	91.49 (14.77)	Total observations = 991/995 (4 missing)	90.78 (14.7)

Source: Pupil enumeration data and pupil BPVS assessments carried out by NatCen in September – October 2022.

Missing data

As a first step, we will explore the extent of missing data on the outcome and pre-treatment covariates descriptively, with cross-tabulations, including counts and percentages in each category.

To better understand the pattern of missing data, we will explore the extent of missingness, and whether there is a pattern in missingness. A 'drop-out' model will be estimated using a logistic regression to assess if there are patterns to missing data. The outcome will be binary, reflecting whether the primary outcome data and any covariates from the primary analysis are missing for each individual at follow-up. This model will include all covariates outlined in the 'imbalance at baseline' section, in addition to a random effect for settings. Missing data for these covariates will be coded up as separate binary variables in the model. The 'drop-out' model will be estimated using the melogit command in Stata 17.

We will follow the protocol for missing data suggested by the EEF (see EEF, 2018). For less than 5% missingness overall from randomisation to final analysis, a complete-case analysis will be employed. For more than 5% missing data overall from baseline assessment to final analysis, our approach will depend on pattern of missingness. If the pattern of missingness is unrelated to the treatment effect (e.g. absence due to child illness, staff changes, or other factors that affected testing but are not related to Hanen LLLI), then missing data will be assumed to be missing completely at random (MCAR) and we will continue with a complete case analysis. Otherwise, the primary analysis will be re-estimated through Multiple Imputation (MI) using Chained Equations (MICE).

MICE will be conducted using the mi suite of commands in Stata 17.

Compliance

The main framework of analysis for this trial is ITT (for more information see the **Error! Reference source not found.**section). However, we will also explore programme effects separately for settings allocated to the treatment group that implemented Hanen LLLI as intended, based on a compliance score.

Based on the EEF's definition, compliance should measure whether the critical ingredients of Hanen LLLI were delivered as intended (EEF, 2019b).

Compliance in this trial is defined at the setting level. A range of items will capture compliance to Hanen, including: 1) Staff attendance at information event; 2) Staff attendance at introductory workshop; 3) Staff attendance at training workshops; 4) Staff attendance at video feedback sessions; 5) Baseline and post-programme visits at the setting (see Table 9).

As shown in Table 9, for a setting to be compliant they need to nominate at least one teacher or room lead to attend the workshops and video feedback sessions (item 3 and 4 in Table 1). It was decided that if there are only TAs attending the workshops and the video feedback

sessions, then the setting would not be considered compliant. This decision was prompted by the fact that TAs alone would not be able to cascade learnings from Hanen. In respect to the attendance at the workshops (item 3), a setting would be compliant if at least one teacher or room lead attended workshops 1, 2, 5 and 3 more workshops (i.e., a total of 6 workshops, including workshops 1, 2, and 5). A staff member can be replaced before workshop 1 only. If a member of staff is replaced after workshop 1, then their setting? would not be deemed compliant. Regarding the video feedback sessions (item 4), a staff member needs to attend at least four of them.

In addition to attendance at the workshops and video feedback sessions, a setting needs to send at least one member of staff to the information event (item 1) and introductory session (item 2). There are no requirements on the role of the staff member that will attend the information event or the introductory session. Settings are also required to have at least one baseline and post-programme visit carried out by the Program Leaders (item 5).

Data for items 2 to 5 will be collected via an attendance register using a template completed by the delivery team. Information on attendance at the information event will be recorded at the MoU summary template. Missing data on any criteria will be scored as zero, i.e., the setting will be considered as non-compliant.

Settings who fulfil all criteria (item 1 through 5) will be deemed compliant. The compliance measure will be a binary indicator (where 1 = compliant; 0 = non-compliant, as defined in Table 1).

Table 9 Compliance measure development

Number	Compliance criterion	Data source	Compliance indicator
1	Attendance at information event	MoU summary template recording attendance	At least 1 member of staff attends
2	Attendance at introductory workshop	Attendance register	At least 1 member of staff per setting attends
3	Attendance of training workshops	Attendance register	At least 1 teacher or room lead attends 6 or more workshops. Workshop 1, 2 and 5 are essential (i.e., if an essential workshop is missed then a setting will not be deemed compliant)*
4	Attendance of video feedback sessions	Attendance register	At least 1 teacher or room lead attends 4 or more video feedback sessions
5	Baseline and post-programme visits	Attendance register	One baseline and post-programme visit per setting are completed

*Staff members (teacher or room lead) can be replaced before workshop 1 i.e. if the replacement is a teacher or room lead and the replacement attends 6 workshops including workshops 1, 2 and 5 it will be considered compliant. If the replacement does not attend workshop 1 they are not compliant. If a setting has nominated only practitioner for the trial who was replaced after workshop 1, then the setting would not be compliant.

In a situation of imperfect compliance (i.e., if some settings are non-compliant), we will undertake a Complier Average Causal Effect (CACE) analysis, by drawing on an

instrumental variable (IV) approach and using a two-stage least squares (2SLS) estimation approach to recover the treatment effect for those who complied with assignment. The first stage estimates if the assignment to Hanen LLLI encourages settings to take up treatment (the first stage regresses treatment assignment on compliance). This provides an estimate of the compliance rate. Results for the first stage will report the correlation between the instrument and the endogenous variable and an F test. The second stage of the IV estimation predicts the outcome using the compliance rate estimated in the first regression by substituting the treatment indicator (assignment to Hanen LLLI) with the compliance rate. The results of this model will answer the research question: 'To what extent does compliance with the Hanen LLLI delivery requirements lead to improved receptive language skills for children?'. This model will be estimated for the primary outcome measure only (i.e., for receptive language skills based on the BPVS).

Intra-cluster correlations (ICCs)

The ICC is a measure of how strongly pupils from the same setting resemble one another in terms of their achieved scores. If the ICC is high, this indicates that pupils from the same setting are likely to score similarly on the assessment (e.g. BPVS-3), due perhaps to sharing a similar learning or home environment. The higher the ICC, the less unique information each new observation from the same cluster contributes to the sample, which will cause the MDES to be higher for a given sample size.

The ICC at analysis stage will be based on the primary outcome measure (i.e., receptive language measured using the BPVS-3) and will be calculated using: (i) the same model as Equation (1) and (ii) a model similar to that documented in Equation (1) but with no covariates accounting for the clustering of pupils in settings (the so-called empty model).

ICCs will be estimated using Stata's estat icc command.

Effect size calculation

We will use the effect sizes (ES) for cluster-randomised trials, as adapted from Hedges (2007):

$$ES = \frac{(\overline{Y_T} - \overline{Y_C})_{adjusted}}{\sqrt{\sigma_s^2 + \sigma_{error}^2}}$$

Where $(\overline{Y_T} - \overline{Y_C})_{adjusted}$ is the mean difference between the intervention and control group adjusted for baseline characteristics, while $\sqrt{\sigma_s^2 + \sigma_{error}^2}$ is an estimate of the population standard deviation.

From the primary outcome model, we will take each group's adjusted mean and variance to calculate the effect size. This variance will be the total variance (across both child and setting levels, without any covariates, as emerging from a 'null' or 'empty' multi-level model with no predictors). A 95% CI for the ES, that takes into account the clustering of children in settings, will also be reported. Effect sizes will be calculated for each of the models estimated.

All ES will be estimated using the eefanalytics Stata package[15].

References

British Picture Vocabulary Scale, Third Edition (BPVS-3): Manual. GL Assessment

Bury, J., Marshall, L., Read, H., Roberts, E., Fletcher, A. and Scandone, B. (2022) Hanen Learning Language and Loving it (LLLI) Evaluation Report, London: EEF

Dimova, S., Scott, M., Scandone, B., Sciarra, A., Rezaian, M., and Bury, J. (2022) Trial evaluation protocol: Learning Language and Loving It, London: EEF

Dong, N. and Maynard, R. (2013) PowerUp!: A Tool for Calculating Minimum Detectable Effect Sizes and Minimum Required Sample Sizes for Experimental and Quasi-Experimental Design Studies, Journal of Research on Educational Effectiveness, 6(1), pp24-67, DOI: 10.1080/19345747.2012.673143

Dunn, L., Dunn, D., Sewell, J., Styles, B., Brzyska, B., Shamsan, Y., Burge, B. (2009) The British Picture Vocabulary Scale Third Edition, Manual. GL Assessment.

Education Endowment Foundation [EEF] (2018) Statistical Analysis Guidance for EEF Evaluations London: EEF

Goodman, R., Meltzer, H., & Bailey, V. (1998). The Strengths and Difficulties Questionnaire: A pilot study on the validity of the self-report version. European child & adolescent psychiatry, 7(3), 125-130.

Goodman. R. (2001). Psychometric properties of the strengths and difficulties questionnaire. Journal of the American Academy of Child and Adolescent Psychiatry, 40 (11), 1337-1345.

Hedges, L. V. (2007) 'Effect Sizes in Cluster-Randomized Designs' Journal of Educational and Behavioral Statistics 32(4): 341–370.

Law, J., Charlton, J., Dockrell, J., Gascoigne, M., McKeane, C. and Theakston, A. (2017) Needs, provision, and intervention for preschool children from socio-economically disadvantaged backgrounds London: Education Endowment Foundation

Piasta, S. B., Justice, L. M., Cabell, S. Q., Wiggins, A. K., Turnbull, K. P. and Curenton, S. M. (2012) 'Impact of professional development on preschool early years practitioners' conversational responsivity and children's linguistic productivity and complexity' Early Childhood Research Quarterly 27 (3):. 387–400.

Renfrew, C. E. (1988). Renfrew action picture test. CE Renfrew.

[15] For more information see EEFANALYTICS: Stata module for Evaluating Educational Interventions using Randomised Controlled Trial Designs (repec.org)