PROJECT TITLE	Hanen: Learning Language and Loving It™ (LLLI)
DEVELOPER (INSTITUTION)	The Hanen Centre
EVALUATOR (INSTITUTION)	NatCen Social Research
PRINCIPAL INVESTIGATOR(S)	Lydia Marshall
PROTOCOL AUTHOR(S)	Lydia Marshall, Robert Wishart, Daniel Phillips
TRIAL DESIGN	Two-arm cluster randomised controlled trial with random allocation at nursery level (Cancelled)
TRIAL TYPE	Efficacy (Cancelled)
PUPIL AGE RANGE AND KEY STAGE	3 to 4 years old, Early Years
NUMBER OF SCHOOLS	147
NUMBER OF PUPILS	2,358[1]
PRIMARY OUTCOME MEASURE AND SOURCE	British Picture Vocabulary Scale (BPVS) (Cancelled)
SECONDARY OUTCOME MEASURE AND SOURCE	Renfrew Action Picture Test (RAPT) (Cancelled)

[1] This is an estimate based on expected numbers of pupils and settings. This will be updated once data has been uploaded for the 2020/21 cohort of pupils.

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0 [original]		N/A[2]
2.0	22/09/2020	COVID-19 redesign for a new cohort of pupils, removing the baseline attainment test and updating power calculations
3.0	07/05/2021	Impact evaluation cancelled based on the methodological risks and ethical challenges of continuing with the RCT in the light of the Covid-19 pandemic

The impact evaluation of Hanen LLLI was cancelled in March 2021 based on the risks and ethical challenges of continuing with the RCT in the light of the Covid-19 pandemic and its implications for programme implementation and data collection. These factors included:

- Sample size and power: The expected MDES at the time the trial was cancelled was 0.23.
- Outcome measurement: There were perceived risks with outcome testing. This includes the ethical concerns about sending external assessors to settings during the COVID-19 pandemic and concerns about burden for school staff. A hybrid model of external assessors and assessments conducted by early years staff was considered, but this also increased risks to the security of trial findings.
- Fidelity: Early Years attendance was markedly lower than usual, particularly in the first half-term of 2021. The delivery partner, evaluator and EEF shared concerns that the impact evaluation findings would not be reflective of Hanen LLLI in a typical academic year, which would make the generalisability and interpretation of findings difficult.

For these reasons the decision was made to focus on the Implementation and Process Evaluation (IPE) only. For further information on the IPE and changes to the IPE see amended Evaluation Protocol[3]. As the impact evaluation was cancelled, the planned analyses as outlined in the SAP will not be conducted but are included below for information. This statistical analysis plan outlines the intended impact evaluation design at the start of the 2020/21 academic year.

[2] Note that whilst an original version of the SAP was prepared and peer-reviewed, it was not published as the COVID-19 lockdown had just been introduced, which had significant implications for the trial design.

[3] Learning Language and Loving It (TM) - The Hanen Program® for Early Childhood Educators | Projects | Education Endowment Foundation | EEF

Table of Contents

Introduction

Learning, Language and Loving It™ - The Hanen Program® for Early Childhood Educators (Hanen LLLI) is a training programme for early years (EY) practitioners to promote social, language and literacy learning in nurseries. It is a continuing professional development (CPD) programme 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.

Hanen LLLI was developed by the Hanen Centre in Canada and is not widely used in the UK. A few Hanen training programmes have been run in the UK over the past two years, focusing mainly on a shortened version of Hanen LLLI, called Teacher Talk. EEF previously funded a **pilot evaluation** of Hanen LLLI to explore its implementation in an English context. This efficacy trial builds on this evidence to explore the impact that Hanen LLLI might have on pupil's language attainment in the early years.

In this evaluation the intervention will be coordinated and delivered by Communicate Speech and Language Therapy Community Interest Company (SLT CIC), a speech and language therapy organisation based in the North West of England. Communicate SLT are Hanen-certified trainers for some of the Hanen programmes but are otherwise not affiliated in any way with The Hanen Centre.

This trial was redesigned in 2020 in response to school closures from Covid-19. The design therefore builds on elements from the evaluation as it was designed for the 2019/20 cohort of pupils, adapted for the 2020/21 cohort of pupils. The 2020/21 cohort will replace the 2019/20 cohort. There are some changes to the planned implementation for 2020/21, to support social distancing. All changes are outlined in the updated trial protocol.

Design overview

Trial design, including number of arms		Two-arm cluster randomised controlled efficacy trial
Unit of randomisation		Nursery
Stratification variables (if applicable)		Geographic region
Primary outcome	variable	Language attainment
	measure (instrument, scale, source)	British Picture Vocabulary Scale (BPVS) raw score (receptive language, scale: 0-168)
Secondary outcome(s)	variable(s)	Language attainment
	measure(s) (instrument, scale, source)	Renfrew Action Picture Test (RAPT) raw score (expressive and receptive language, scale: 0-80)
Baseline for primary outcome	variable	None[4]
	measure (instrument, scale, source)	None
Baseline for secondary outcome	variable	None
	measure (instrument, scale, source)	None

Sample size calculations overview

Original design

Communicate SLT aimed to recruit 147 nurseries in order to allow for any drop-out over the summer break and over the course of the evaluation. In total, Communicate SLT approached 861 settings, recruiting 147 to the Hanen LLLI trial.[5] In August 2019, 73 settings were randomly

[4] The original trial design included collected the British Picture Vocabulary Scale (BPVS) which would have been used as a baseline measure for both the primary and secondary outcomes. However, based on the expected burden on schools and restrictions on school visitors, baseline data collection was not conducted for the 2020/21 cohort of pupils.

[5] Three settings dropped out prior to randomisation and were replaced with settings on a waiting list.

allocated to a treatment condition in which nurseries receive the Hanen LLLI intervention while the remaining 74 were randomly allocated to a 'business as usual' control condition.

Since randomisation, sixteen schools dropped out of the evaluation, fifteen prior to baseline testing (seven intervention; eight control) and one intervention school after baseline testing. Although these settings had already been randomised, they (and the developers) were blind to their allocation at the point at which they dropped out of the evaluation.[6] The intention-to-treat sample therefore consists of 65 intervention settings and 66 control settings.

For education programmes, the variance explained by pre-test scores can be relatively high if pre-test scores are used in adjusted analysis (Bloom et al, 2007). This increases statistical power, making it easier to detect small effects. Language attainment was collected at baseline for the 2019/20 cohort of pupils. The pre- and post-test correlations were informed by Torgerson and Torgerson (2013), who reference a pre-test post-test correlation of 0.70. However, the power calculations have been conducted using R-squared, informed by NatCen and EEF's prior experience in evaluations with similar age groups (see footnotes 16~~163~~, 17~~174~~ and 18~~185~~). Therefore, we assumed a pre- post-test correlation of 0.60 at pupil level and 0.36 at school level. Nursery-level intra-cluster correlations (ICCs) were informed by the evaluation of Family Skills (Husain et al., 2018) and Demack (2019). The ICC is therefore estimated to be 0.15 at school-level. At randomisation, the MDES was estimated as 0.18 for the primary analysis and 0.29 for the sub-group analysis. For reference, the power calculations for the original design are available in Appendix A.

Re-design

Although similar, the approach for the re-design has several differences from the original design. One of these changes is the decision not to conduct a baseline attainment test for the 2020/21 cohort of pupils. This decision was taken because of the burden it would place on schools and pupils in the first term back (for many schools and pupils) since schools closures in 2019/20. There is therefore no measure of baseline attainment in the model.

A further consideration for the trial is how many nurseries are still willing to take part in the evaluation, given that the original evaluation was due to end in 2020. At the time of writing, 51 intervention nurseries had confirmed they would take part in 2020/21. One intervention nursery had closed, two had dropped out in the previous academic year, five nurseries indicated they would not continue with the evaluation and three nurseries were considering whether to continue or not.

By comparison, 44 control nurseries had confirmed they would continue, five where considering whether to continue or not and four nurseries indicated they would not continue with the evaluation. A further 13 nurseries had not yet responded.

Our power calculations assume that all nurseries that are currently considering whether to continue or not will take part in the evaluation. However, we have relaxed the maximum

[6] These schools are not considered in the intention-to-treat sample. These cases will therefore be included in estimates of attrition.

number of pupils per setting restriction in the re-design. Increased pupil numbers (maximum of 20 per setting) may yield additional power.

Table 1: Minimum Detectable Effect Size Calculations

		Protocol		Randomisation	
		OVERALL	EYPP[7]	OVERALL	EYPP
Minimum Detectable Effect Size (MDES)		0.22	0.37	0.25	0.42
Pre-test/ post-test correlations[8]	level 1 (pupil)[9]	0.00	0.00	0.00	0.00
	level 2 (school)[10]	0.00	0.00	0.00	0.00
Intracluster correlations (ICCs)	level 2 (school)	0.15	0.15	0.15	0.15
Alpha		0.05	0.05	0.05	0.05
Power		0.80	0.8	0.8	0.8
One-sided or two-sided?		2	2	2	2
Average cluster size		18	2	18	2
Number of schools	intervention	65	65	54	54
	control	66	66	49	49
	total	131	131	103	103
Number of pupils	intervention	1,170	130*	972	108*
	control	1,188	132*	792	98*
	total	2,358	262*	1,764	206*

* Proportion of FSM pupils anticipated to be national average for age group (nursery and primary school age) of 14.2%, as in DfE (2018c). We note that EEF protocols usually include sub-group analysis by Free School Meal (FSM) status. However, we use Early Years Pupil Premium (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 are also eligible for FSM.

The calculations were undertaken using PowerUp! (Dong and Maynard, 2013) and indicate that this study is powered to detect an effect of 0.26 standard deviations based on the

[7] We note that EEF protocols usually include sub-group analysis by Free School Meal (FSM) status. However, we use Early Years Pupil Premium (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.

[8] Our sample size calculations include estimates of the proportion of variance explained through the included covariates at each of these levels, also known as R-squared. The R-squared values have been estimated by squaring the pre-test post-test correlation.

[9] An R-squared value of 0.36 at level 1 is used in the power calculations.

[10] An R-squared value of 0.18 at level 2 is used in the power calculations.

randomised sample. The Early Years Pupil Premium (EYPP) subgroup analysis will be powered to detect an effect of 0.42 standard deviations. We also present a scenario, assuming that only those that have confirmed they will continue are included. We assume similar numbers of pupils per setting. In this case, the MDES would be 0.27.

The calculations do not account for further attrition over the course of the evaluation. A full CONSORT diagram, detailing attrition and its causes at different stages, will be provided in the evaluation report.

Analysis

Primary outcome analysis

The evaluation of Hanen LLLI aims to estimate the impact of the programme on the language attainment of three to four-year olds in England and how it differs by EYPP eligibility, using an intention-to-treat approach. The trial is designed as a two-armed cluster randomised efficacy trial with pupils (level one) clustered within nurseries (level two).

To account for the clustering of pupils within nurseries, the impact will be estimated using a two-level linear regression model. The BPVS raw score at follow-up will be regressed on a binary indicator of treatment allocation, and geographic region (the randomisation strata). School-level random effects will be included in the model by allowing the intercept to vary randomly across schools.

The model will take the following form:

$$Language\ attainment_{ij} = \beta_0 + \beta_1 Intervention_j + \beta_2 Region + u_j + e_{ij}$$

Where pupils (i) are nested within schools (j). The intervention effect is estimated by β_1, whilst u_j represents the school random effect and e_{ij} represents the individual error term. In line with EEF guidance (EEF, 2018) other covariates will not be considered at this stage but will be explored through additional analyses (see below). The analysis will be conducted in Stata 16 SE using the **mixed** command.

Secondary outcome analysis

The secondary outcome analysis will assess the impact of Hanen LLLI using the Renfrew Action Picture Test (RAPT), on an intention-to-treat basis. The analytical approach is identical to that set out in the primary analysis, using a multilevel linear regression model, with pupils nested within schools. The raw RAPT score at follow-up will be regressed on a binary indicator of treatment allocation and geographic region (the randomisation strata). Using a different baseline measure of attainment may lower the pre- post- test correlation, reducing the power of the secondary analysis. School-level random effects will be included to account for the variance at school level.

Subgroup analyses

A single sub-group analysis will explore if Hanen LLLI has a differential impact on pupils eligible for Pupil Premium.[11] EYPP provides schools with additional funding for all three to four-year-olds from low-income families. All EYPP pupils are also eligible for FSM.

The analytical approach is similar to the primary and secondary analyses. A multilevel linear regression model will be employed, where the BPVS raw score will be regressed on a binary indicator of treatment allocation, geographic region (the randomisation strata), an indicator for EYPP eligibility, an interaction term for treatment allocation and EYPP eligibility and a random effect for schools.

$$\boldsymbol{Language\ attainment_{ij}} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 Intervention_j} + \boldsymbol{\beta_2 EYPP_{ij}} + \boldsymbol{\beta_3 intervention * EYPP_{ij}} + \boldsymbol{\beta_4 Region} + \boldsymbol{u_j} + \boldsymbol{e_{ij}}$$

If this interaction term coefficient is statistically significant, an additional model, taking the same form as the primary analysis model will be estimated, using solely the sample of pupils eligible for EYPP.

$$Language\ attainment\ (EYPP\ only)_{ij} = \beta_0 + \beta_1 Intervention_j + \beta_2 Region + u_j + e_{ij}$$

Additional analyses

The primary analysis uses multi-level modelling, a technique based on strong assumptions (such as normally distributed residuals within each cluster). As a sensitivity test, a single-level OLS regression model, using Huber-White cluster-robust standard errors will be estimated. The model will use the same covariates outlined for the primary analysis. This will be estimated in Stata 16 SE using the **reg** command. The model will take the form:

$$Language\ attainment_{ij} = \beta_0 + \beta_1 Intervention_j + \beta_2 Region + e_{ij}$$

Longitudinal follow-up analyses

- Follow-up analysis can be conducted using data from the National Pupil Database (NPD).[12] Children are assessed for the Early Years Foundation Stage Profile (EYFSP) at the end of the school year they turn five. The EYFSP contains three measures that may be suitable for assessing the longer-term impacts of the intervention: G01: Communication and Language – Listening
- G02: Communication and Language – Understanding
- G03: Communication and Language – Speaking

These components can be used to assess expressive and receptive language, corresponding to the intervention's logic model. Therefore, we propose combining these measures into an

[11] We note that EEF protocols usually include sub-group analysis by Free School Meal (FSM) status. However, we use Early Years Pupil Premium (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.

[12] Longitudinal follow-up analyses are not funded as part of this evaluation.

index. However, this outcome could be revised in line with any future revision to the logic model.

Imbalance at baseline

Imbalance at baseline in pupil or setting characteristics could occur by random chance. Individual characteristics, such as EYPP status, and differences between settings, could impact on language attainment.

Pupil and setting characteristics will be analysed for the 'as analysed' and 'as randomised' samples to explore potential imbalance arising from the randomisation or subsequent attrition.

At pupil level, the following baseline comparisons will be presented:

- Early Years Pupil Premium status

At setting level, the comparison will cover:

- Prior language attainment, measured by the 2019/20 baseline British Picture Vocabulary Scale (collected for the original evaluation design)
- Setting's Ofsted rating
- Percentage of pupils eligible for EYPP in the setting
- Setting type (School based nursery or maintained nursery)
- Whether the setting is in an urban or rural area

Potential imbalance for categorical variables will be explored by conducting cross-tabulations, including counts and percentages in each category, and tested with Chi-Square tests. For continuous variables, we will report descriptive statistics (n, mean, standard deviation, range, median and effect sizes) by group allocation and use paired sample t-tests. We will report differences in baseline characteristics as Hedge's g effect sizes. An effect size of greater than 0.05 will be considered as an indication of possible imbalance.

If imbalances are indicated, a sensitivity analysis will be estimated. This model will include the same covariates as the primary analysis, a random intercept for settings and all unbalanced variables (i.e. where Hedge's g is greater than 0.05).

Missing data

The extent of missing data on the outcome and pre-treatment covariates will be analysed descriptively, with cross-tabulations, including counts and percentages in each category. Differences in the extent of missing data by allocation will be tested with Chi-Square tests

Guidance from the EEF (2019) suggests that attrition in early-years trials is likely to be approximately ten percent at setting level and over twenty percent at individual level, when a trial is conducted over a single academic year. This can occur for several reasons, such as pupils moving to different settings and absence at baseline or follow-up testing.

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 schools. 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 16 SE.

The primary analysis will be re-estimated through Multiple Imputation (MI) using Chained Equations (MICE) if:

- Greater than five percent of pupils are missing data on the outcome or covariates included in the primary analysis; and,
- The 'drop-out' model finds statistically significant associates between covariates and the outcome of the 'drop-out' model.

The MI analysis will draw on all available information from the trial data. This includes:

- BPVS attainment at baseline for the 2019/20 cohort, aggregated at nursery level
- Intervention allocation
- BPVS and RAPT attainment at follow-up
- EYPP status

Multiple imputation will be conducted using the **mi** suite of commands in Stata 16 SE. The first 100 iterations of the imputation will not be used ('burn in'). This ensures that iterations have converged to a stationary distribution. In total, 100 imputed datasets will be created. The analysis will then be estimated using the **mi estimate: mixed** command in Stata 16 SE.

Compliance

To be eligible to take part in the intervention, a nursery must be able and agree to release a minimum of 50% of its eligible staff for the training. 'Staff' includes nursery teacher(s), plus nursery nurses, teaching assistants (TAs), and/or other more senior staff.

A measure of compliance will be constructed according to staff attendance at Hanen LLLI training. The protocol indicates that the Hanen LLLI programme includes eight training sessions, six individual feedback sessions and an orientation session. Feedback sessions focus on guided reflection, where Program Leaders provide feedback on videotaped interactions between EY practitioners and children. In addition, there are also two video sessions; one at the start of the programme and one at the end (Marshall et al, 2018), which are also included in the compliance measure.

Attendance at training will be captured via a register designed by NatCen and completed by Communicate SLT.

Compliance will be identified as a continuous variable at school level and constructed as follows:

- Nurseries will be given one point for each of the following sessions attended by a teacher[13]:

[13] Nurseries will be given a recommendation of signing up two-thirds of staff working with 3 to 4-year-olds to take part in Hanen LLLI, but the minimum requirement will be that at least 50 per cent of eligible practitioners including a teacher should be able to participate.

- Orientation session (up to 1 point)
- Eight workshops (up to 8 points)
- Six video sessions (up to 6 points)
- The baseline video session and post video session (up to 2 points)

- Two further points will be awarded to any nursery where the head-teacher (or equivalent, such as a deputy head or member of the senior leadership team) attends at least one training workshop.
- Nurseries will be awarded further points for the proportion of sessions (comprising all sessions outlined above) attended by one additional[14] staff member who had been nominated to attend training, up to a total of four points. The proportion of sessions attended on average by the additional nominated staff member will be used to calculate additional points. For example, if average additional staff attendance for a particular nursery is 25 per cent or more, the nursery will be awarded an additional one point (≥ 50%, 2 points; ≥ 75%, 3 points; 100%, 4 points;).

The total score will then be summed to produce a scale of compliance with a possible range of $0 \leq Comply_A \leq 23$ which will then re-scaled to have a range of $0 \leq Comply_A \leq 1$. As compliance is defined through attendance to training, it is anticipated that there will be one-sided non-compliance.

Compliance will be analysed using an Instrumental Variable (IV) approach where compliance is instrumented using allocation status (Angrist and Imbens, 1995). The instrumental variable regression will use two-stage least squares (2SLS), where the first stage equation estimates

$$Comply_j = \propto + \beta_1 Treat_j + \varepsilon_{ij}$$

The predicted values from the first stage equation, $\widehat{Comply_j}$, will then be used in the estimation of the second stage equation, as follows:

$$Y_{ij} = \propto + \beta_1 Treat_j + \beta_2 \widehat{Comply_j} + \beta_3 Region_{ijk} + \omega_{ij}$$

IV regression will be conducted in Stata 16 using the **ivregress** command and cluster robust standard errors estimated with the **cluster** option. Tests for endogeneity will be conducted to test if treatment allocation is a suitable instrument (Wooldridge, 1995) and following EEF guidance (EEF, 2018) the F-statistics and p-value will be reported. If compliance is exogenous, the analysis will be conducted using multi-level linear regression.

Intra-cluster correlations (ICCs)

The intra-cluster correlations (ICC) will be estimated separately for the baseline and follow-up attainment test. The ICCs will be estimated from the variance components of a multilevel linear regression model containing treatment allocation as the sole covariate and a random effect for schools. The ICC; $\boldsymbol{\rho}$, will be estimated with the post-estimation command estat icc in Stata 16 SE, using the following formula:

[14] 'Additional' means any staff other than the nominated teacher, or head-teacher/equivalent.

$$\rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2} = \frac{\sigma_B^2}{\sigma_T^2}$$

Where $\boldsymbol{\sigma_B^2}$ is the between-school variance, $\boldsymbol{\sigma_W^2}$ is the within-school variance and $\boldsymbol{\sigma_T^2}$ is the total variance.

Effect size calculation

The Hedge's g effect size will be estimated following Hedge's (2007) formulae for the effect size d_t for designs with unequal sample sizes. This ensures that clustering is accounted for in the estimation of the treatment effect and standard errors. The effect size, g_t is estimated as follows:

$$g_t = J \times \left(\frac{\bar{Y}^T_{\blacksquare\blacksquare} - \bar{Y}^C_{\blacksquare\blacksquare}}{S_T}\right)\sqrt{1 - \rho\left(\frac{(N - n_u^T m^T - n_u^C m^C) + n_u^T + n_u^C - 2}{N - 2}\right)}$$

Where:

- $\boldsymbol{\bar{Y}^T_{\blacksquare\blacksquare}}$ and $\boldsymbol{\bar{Y}^C_{\blacksquare\blacksquare}}$ are the grand means of the treatment and control groups
- $\boldsymbol{\rho}$ is the intra-cluster correlation
- N is the total number of pupils
- M is the total number of schools, divided between the intervention group $\boldsymbol{m^T}$ and the control group $\boldsymbol{m^C}$

The remaining terms are calculated as follows:

The correction factor J is defined as:

$$J = 1 - \left(\frac{3}{4(N^T + N^C - 2) - 1}\right)$$

The pooled standard deviation, S_T is defined as:

$$S_T = \sqrt{\frac{\sum_{i=1}^{m^T}\sum_{j=1}^{n_i^T}(Y_{ij}^T - \bar{Y}^T_{\blacksquare\blacksquare})^2 + \sum_{i=1}^{m^C}\sum_{j=1}^{n_i^C}(Y_{ij}^C - \bar{Y}^C_{\blacksquare\blacksquare})^2}{N - 2}}$$

And the term n_u^C is defined in the same way as n_u^T:

$$n_u^T = \frac{(N^T)^2 - \sum_{i=1}^{m^T}(n_i^T)^2}{N^T(m^T - 1)}$$

The variance term is calculated as follows:

$$V\{g_t\} = \left(\frac{N_T + N_C}{N_T N_C}\right)(1 + (\tilde{n} - 1)\rho) + \frac{[(N-2)(1-\rho)^2 + A\rho^2 + 2B\rho(1-\rho)]\delta_g^2}{2(N-2)[(N-2) - \rho(N-2-B)]}$$

Where:

$$\tilde{n} = \frac{N^C \sum_{i=1}^{m^T}(n_i^T)^2}{N^T N} + \frac{N^T \sum_{i=1}^{m^C}(n_i^C)^2}{N^C N}$$

And:

$$A = A^T + A^C$$

Where A^C is calculated in the same way as A^T using control group sample sizes:

$$A^T = \frac{(N^T)^2 \sum_{i=1}^{m^T}(n_i^T)^2 + \left(\sum_{i=1}^{m^T}(n_i^T)^2\right)^2 - 2N^T \sum_{i=1}^{m^T}(n_i^T)^3}{(N^T)^2}$$

B can be calculated as follows:

$$B = n_u^T(m^T - 1) + n_u^C(m^C - 1)$$

Confidence intervals:

Finally, confidence intervals for a two-tailed test are calculated as follows:

$$\delta_g - c_{\frac{\alpha}{2}} v_g \leq \delta_g \leq \delta_g + c_{\frac{\alpha}{2}} v_g$$

Where $c_{\frac{\alpha}{2}}$ is the critical value for a type one error rate, α, for a given sample size.

References

Angrist, J. and Imbens, G. (1995) `Two-stage least squares estimation of average causal effects in models with variable treatment intensity.' American Statistical Association, 90(430), pp431-442

Bloom, H.S., Richburg-Hayes, L. and Black, A.R. (2007) 'Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions' Educational Evaluation and Policy Analysis 29 (1): 30–59.

DeMack, S. (2019) 'Does the Classroom Level Matter in the Design of Educational Trials? A Theoretical and Empirical Review'. EEF Research Paper No. 003 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

EEF (2018) `Statistical analysis guidance for EEF evaluations', London: EEF.

EEF (2019) `Lessons learnt from EEF Early Years trials: Recommendations for evaluators', London: EEF.

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

Husain, R., Wishart, R., Marshall, L., Frankenberg, S., Bussard, L., Chidley, S., Hudson, R., Vojtkova, M and Morris, S. (2018) `Family Skills Evaluation report and executive summary' London: EEF.

Marshall, L., Phillips, D. and Wishart, R. (2018) `Trial evaluation protocol: Learning Language and Loving It™', London: EEF.

Torgerson, C. and Torgerson, D. (2013) `Randomised trials in education: An introductory handbook' London: EEF

Wooldridge, J. M. 1995. `Score diagnostics for linear models estimated by two stage least squares'. In `Advances in Econometrics and Quantitative Economics: Essays in Honor of Professor C. R. Rao', ed. G. S. Maddala, P. C. B. Phillips, and T. N. Srinivasan, 66-87. Oxford: Blackwell.

Appendix A: Power calculations

Table 2 Minimum detectable effect size (MDES) calculations; original design

		Protocol		Randomisation	
		OVERALL	EYPP[15]	OVERALL	EYPP
Minimum Detectable Effect Size (MDES)		0.19	0.30	0.18	0.29
Pre-test/ post-test correlations[16]	level 1 (pupil)[17]	0.60	0.60	0.60	0.60
	level 2 (school)[18]	0.36	0.36	0.36	0.36
Intracluster correlations (ICCs)	level 2 (school)	0.15	0.15	0.15	0.15
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		17	2	16	2
Number of schools	intervention	70	70	73	73
	control	70	70	74	74
	total	140	140	147	147
Number of pupils	intervention	1,190	169*	1,168	166*
	control	1,190	169*	1,184	168*
	total	2,380	338*	2,352	334*

[15] We note that EEF protocols usually include sub-group analysis by Free School Meal (FSM) status. However, we use Early Years Pupil Premium (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.

[16] Our sample size calculations include estimates of the proportion of variance explained through the included covariates at each of these levels, also known as R-squared. The R-squared values have been estimated by squaring the pre-test post-test correlation.

[17] An R-squared value of 0.36 at level 1 is used in the power calculations.

[18] An R-squared value of 0.18 at level 2 is used in the power calculations.