# *Headsprout Early Reading® in Special Schools (HERiSS): a randomised control trial*

# Statistical Analysis Plan

Evaluator (institution): The University of Warwick

Principal investigator(s): Dr Samantha Flynn & Dr Louise Denne

| | |
|---|---|
| **PROJECT TITLE**[1] | Headsprout Early Reading® in Special Schools (HERiSS) – A randomised control trial. |
| **DEVELOPER (INSTITUTION)** | Bangor University |
| **EVALUATOR (INSTITUTION)** | The University of Warwick |
| **PRINCIPAL INVESTIGATOR(S)** | Dr Samantha Flynn, & Dr Louise Denne |
| **PROTOCOL AUTHOR(S)** | Dr Louise Denne, Dr Samantha Flynn, Dr Paul Thompson |
| **TRIAL DESIGN** | Two-arm cluster randomised controlled trial with random allocation at school level |
| **TRIAL TYPE** | Efficacy |
| **PUPIL AGE RANGE AND KEY STAGE** | 5 – 11 yrs (KS 1 & 2) |
| **NUMBER OF SCHOOLS** | 99 (at design stage) |
| | 55 (at randomisation stage) |
| **NUMBER OF PUPILS** | 891 (at design stage) |
| | 378 (at randomisation stage) |
| **PRIMARY OUTCOME MEASURE AND SOURCE** | The composite score of the Dynamic Indicators of Basic Early Literacy Skills® 8th edition, Year 2 version (DIBELS®) will be used as the primary outcome to gather data on pupils' reading skills. (Pupil testing) |
| **SECONDARY OUTCOME MEASURES, MODERATORS AND SOURCE** | All DIBELS® components will be used as secondary outcome measures: Letter Naming, Phonemic Segmentation, Nonsense Word, Word Reading, Oral Reading. (Pupil testing)<br><br>Reading self-concept scale (adapted from Chapman and Tunmer, 1995) (Pupil questionnaire)<br><br>Pupil characteristics: school year, first language, receipt of school meals, primary need, free school meals (School or National Pupil Data (NPD)) |

## SAP version history

| VERSION | DATE | REASON FOR REVISION |
|---|---|---|
| 1.0 | 17.11.2021 | N/A |
| **1.1** | 06.05.2022 | Revisions based on external peer review |

## Table of contents


SAP version history ..... 2
Table of contents ..... 2
Introduction ..... 2
Design overview ..... 3
Sample size calculations overview ..... 4
Analysis ..... 7
- Primary outcome analysis ..... 7
- Secondary outcome analysis ..... 8
- Subgroup analyses ..... 9
- Additional analyses ..... 10
- Longitudinal follow-up analyses ..... 10
- Imbalance at baseline ..... 10
- Missing data ..... 10
- Compliance ..... 10
- Intra-cluster correlations (ICCs) ..... 12
- Effect size calculation ..... 12


## Introduction

This evaluation aims to test the effectiveness of the Headsprout® Early Reading (HER®) reading programme that was designed to teach beginning readers skills and strategies using phonics instruction. The intervention will be delivered in special schools for pupils aged between 5 and 11, Key Stages 1 & 2 (KS1 & KS2) who are lacking in prior reading skills. Schools were informed during the recruitment process that eligible pupils would be beginning readers who met the eligibility criteria. These criteria included various pre-requisite learning skills, and a short passage of text to read. If pupils were able to read the passage independently, this would indicate some decoding skills had already been acquired, and they would not be deemed eligible for the study. Approximately 480 pupils across 60 special schools in the United Kingdom will be recruited and randomly assigned to either intervention or teaching as usual group. This efficacy trial will be a two-armed parallel cluster RCT

(cRCT). Schools will be randomly assigned to either intervention or 'teaching as usual' (control) groups.. Schools receiving the intervention and associated training will implement HER® with a group of eligible pupils including access to an accompanying support manual for teaching staff supporting children using HER® that has been developed by the delivery team. For the schools receiving the intervention, this will replace any phonics instruction that would otherwise be implemented with the participating pupils.

Our primary research question will evaluate whether children participating in the Headsprout® Early Reading (HER®) intervention differ significantly from the control group after adjustment for baseline measurements of primary outcome and stratification variable, school-size. From this analysis, we will estimate the effect size of intervention and assess the robustness of the trial. In addition, we will account for additional covariates (Age (school year); whether child is in receipt of free school meals; type of primary need (e.g. Autism, Speech, Language and Communication needs, Social, Emotional and Mental Health needs); and English as a first language), compliance (6-item checklist, including for example, All Teachers/TA's supporting pupils through HERiSS have been through the digital training; and at least one member of staff attending the two webinars). In addition, we will assess the sensitivity of findings under different assumptions with respect to missing data.

Our secondary research question will examine whether the control and intervention groups differ on a number of sub-tests from the Dynamic Indicators of Basic Early Literacy Skills® 8th edition (DIBELS®). Analyses on these outcomes will follow the same statistical approach, detailed above, for primary outcomes.

Further statistical analyses will explore whether group differences in intervention and control groups are moderated by covariatesor depends on covariates..

## Design overview

| | | |
|---|---|---|
| **Trial design, including number of arms** | | Two-arm parallel cluster randomised controlled trial |
| **Unit of randomisation** | | School level |
| **Stratification variables** | | School size (>=70 pupils vs <70 pupils)[2] |
| **Primary outcome** | variable | Pupil's reading skills (Pupils' ability to sound out words – not comprehension) |
| | measure (instrument, scale, source) | A composite score derived from DIBELS® (Pupil testing) |
| **Secondary outcome(s)** | variable(s) | Components of reading fluency (e.g., letter naming fluency, phonemic segmentation fluency, nonsense word fluency, word reading fluency, and oral reading fluency.)<br><br>Pupils' reading self-concept (including reading for pleasure/spontaneous reading) |

[2] The cut-off of 70 pupils as large vs small schools size was derived from SEN data tables provided by the DfE in 2018. Available from: https://www.gov.uk/government/statistics/special-educational-needs-in-england-january-2018 (7)

| | | |
|---|---|---|
| | measure(s) (Instrument, scale, source) | Reading self-concept scale (Pupil questionnaire)<br><br>DIBELs component analysis (Pupil testing) |
| **moderators** | variable(s) | Age (school year), Receipt of free school meals (FSM), type of primary need, English as a first language (EFL)<br><br>Teaching as usual |
| **Baseline for primary outcome** | **variable** | Pupils' reading skills (Pupils' ability to sound out words – not comprehension) |
| | measure (instrument, scale, source) | DIBELS® composite score [] [3] |
| **Baseline for secondary outcome** | **variable** | Pupils' reading self-concept (including reading for pleasure/spontaneous reading)<br><br>Components of reading fluency (e.g. letter naming fluency, phonemic segmentation fluency, nonsense word fluency, oral reading fluency, and word reading fluency) |
| | moderators | Age (school year), Receipt of free school meals (FSM), Type of primary need, English as a first language (EFL) |
| | measure (instrument, scale, source) | Reading self-concept scale (Pupil questionnaire)<br><br>DIBELS® component analysis (Pupil testing)<br><br>Pupil characteristics: school year, first language, receipt of school meals, and primary need (All from School where possible, or NPD, if FSM info unavailable)<br><br>A teaching as usual questionnaire for all schools (Teacher questionnaire) |

[3] The DIBELS composite measure is derived following the DIBELS GUIDE (https://dibels.uoregon.edu/sites/dibels1.uoregon.edu/files/2021-06/dibels_8_composite_score_calculation_guide_supplement_072020.pdf)

## Sample size calculations overview

Sample size calculations were conducted using R version 4.1.2 (2021-11-01) and PowerUpR version 1.1.0 (1,2,3).

Working with the delivery team and previous experience from testing in special schools, we conservatively estimated that 15% of all pupils in a special school may be eligible for HER®, and so we estimated that we would be able to recruit between 5 and 15 primary age pupils

from most special schools in selected regions in England(North West and West Midlands). Prior to assessments moving online due to the Covid -19 pandemic, these regions were initially selected for logistical reasons as it was deemed feasible for the delivery team to support schools across these regions, and for the evaluation team to conduct assessments across these regions. We anticipated that this number of pupils will be manageable for two staff (a trained teacher and Teaching Assistant (TAs)) per school delivering HER®.

We will aim to randomise 55 schools (28 schools in the control arm and 27 schools in the treatment arm of the study). Calculations are based on 80% power, a (two-sided) significance level of 5%, with (following pupil-level attrition, *see below*) an average of 2.33 pupils within each class, and an average of 3 classes within each school at follow-up[3]. Although this is a relative unknown in the case of special schools, we have cautiously accounted for an intracluster correlation coefficient (ICC) as high as .40 at the class level and .10 at the school level, in accordance with guidance (4, 5). Demack (2019)(5) provides a theoretical and empirical review of three level cluster randomised trials in education with the purpose of exploring the implications of ignoring classroom level clustering. There are few special school education trials, and none with sufficient sample size to obtain accurate estimates of appropriate ICCs. The current study would be the largest education trial in a special school context to date. Therefore, we assume estimates of mean ICCs (Primary) recommended from the Demack review's discussion as conservative moderated by expert knowledge of the special school environment by both delivery and evaluation teams. We have allowed for an estimated correlation of pre and post-intervention scores (both on the same DIBELS® composite test) on the outcome of .5, representing a moderate correlation between the two time-points (6). The power calculation results in a minimum detectable effect size of .412.

For the free school meals (FSM) sub-group analysis, we estimate a minimum detectable effect size of .465; based upon the above assumptions, and official statistics showing 43% of FSM pupils in state-funded special schools (7; see https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics).

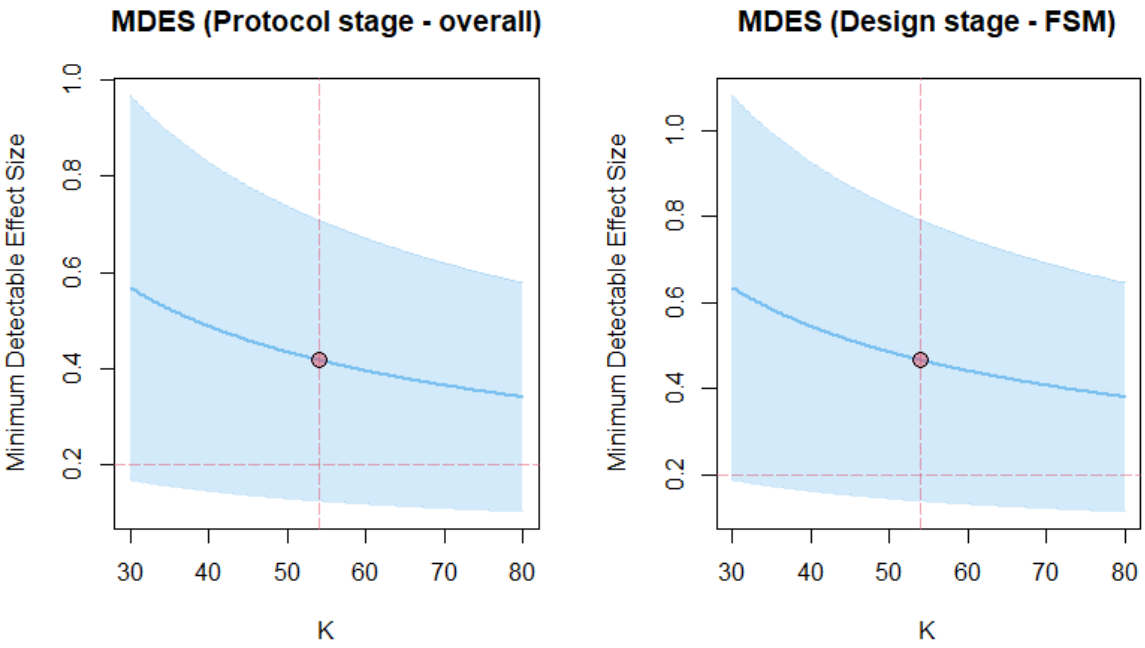
We have allowed for a 10% attrition rate for schools (for both groups), and one pupil per school lost to attrition (11.1% pupil level attrition from average of 8 pupils initially [slightly lower than the originally expected 10] recruited per retained school) in retained schools across the trial. Our drop-out rate is considered as a somewhat conservative estimate, as whilst both lower and higher rates of attrition have been observed in extant studies of the HER® intervention (8,9) this is a relative unknown with respect to large-scale studies in a special school setting. We have prepared the above sample size calculations and the remainder of this protocol under the assumption that 60 schools will be recruited (with 54 schools retained) and 8 pupils recruited within each school (with an average of 7 pupils retained) to achieve a sample of 378 pupils (after attrition at both the school and pupil level) and have budgeted as such.

[3] There are no official published figures on the number of pupils from special schools that schools can put forward for studies. The delivery and evaluation teams made the estimate that schools would have at least 5 eligible pupils to participate and that they could put forward and up to 15 pupils. We then conservatively reduced this number at randomisation stage to 7 based on recruitment estimates.

**Table 2: Sample size calculations (following attrition)**

| | | Design stage | | Protocol/Randomisation stage | |
|---|---|---|---|---|---|
| | | OVERALL | FSM | Overall | FSM |
| Minimum Detectable Effect Size (MDES) | | 0.298 | 0.327 | .412 | .461 |
| Pre-test/ post-test correlations | level 1 (pupil) | 0.50 | 0.50 | 0.50 | 0.50 |
| | | | | | |
| | | | | | |
| Intracluster correlations (ICCs) | level 2 (class) | 0.40 | 0.40 | .40 | .40 |
| | level 3 (school) | 0.10 | 0.10 | .10 | .10 |
| Alpha | | 0.05 | 0.05 | 0.05 | 0.05 |
| Power | | 0.8 | 0.8 | 0.8 | 0.8 |
| One-sided or two-sided? | | Two-sided | Two-sided | Two-sided | Two-sided |
| Average cluster size | | 9 | 3.89 | 7 | 3.02 |
| Number of schools | Intervention | 49 | 49 | 27 | 27 |
| | Control | 50 | 50 | 28 | 28 |
| | **Total** | 99 | 99 | 55 | 55 |
| Number of pupils | Intervention | 441 | 191 | 189 | 82 |
| | Control | 450 | 194 | 196 | 85 |
| | **Total** | 891 | 385 | 378 | 167 |

**Figure 1: Plot showing the minimum detectable effect size as a function of number of schools (K).**




## Analysis

An intention-to-treat (ITT) approach will be used (including all randomised schools and pupils in the analysis), specifying random intercepts to account for the between-classroom and between-school variability, and reporting standard errors of the parameter estimates. ICCs (at both the classroom and school level) will be calculated for the null model (i.e. that without covariates) at post-test.

### *Primary outcome analysis*

The primary analysis will examine mean follow-up DIBELS® composite scores, adjusting for the respective baseline measure school size, and with the covariate of interest specified as a dichotomous treatment/control variable.

The analysis for the primary outcome measure will use a linear mixed model (LMM), given that the measure is continuous distributed. Some amount of skew in the raw outcome measure can be permitted, provided that the assumption of the normality of residuals is satisfied (assessed by looking at the residual diagnostic plots). The range of possible scores and calculation of the composite, indicates that a LMM should be the most appropriate model in this instance. Model assumptions will be checked at the analysis stage and, if necessary, a generalised linear mixed model with appropriate link function can be changed to match the analysis proposed for the secondary outcomes permitting analysis of sufficiently skewed data with heterogeneous residuals (10). Nested model comparison will be based on likelihood ratio tests (Chi square) and both Bayesian and Akaike's Information Criteria (BIC and AIC respectively), with lower indices indicating the preferred model (14). All analyses will be conducted in R (version 4.0.3 - 2020-10-10), using the R packages: Tidyverse, lme4, ordinal, lmerTest (1,11,12,13).

Due to the number of class clusters within school being relatively small (14,22), we will adopt a two-step process to explore issues with over or under-fitting using a sensitivity analysis. In the first step, we will include a two-level design (pupil - level 1; school – level 2), then a second analysis fill follow with the inclusion of the third level (pupil – level 1; class – level 2; and school – level 3).

***Step 1 model (two-level design):***

$$\boldsymbol{L1}: Y_{ik} = \beta_{0k} + \beta_{1k}X1_{ik} + r_{ik},\ r_{ik} \sim N\left(0, \sigma^2{}_{|X}\right)$$

$$\boldsymbol{L2}: \beta_{0k} = \gamma_{00} + \gamma_{20}X2_k + \gamma_{30}INT_k + \mu_{0k}, \qquad \mu_{0k} \sim N(0, \tau_2^2)$$

$$\beta_{1k} = \gamma_{10}$$

where, $Y$ are the DIBELS® composite scores; $X1$ are the baseline DIBELS® composite scores; $INT$ is the treatment/control variable; $X2$ are indicator of school size (strata variable). $\mu_{0k}$ is the random intercept term for school.

***Step 2 model (three level design):***

$$\boldsymbol{L1}: Y_{ijk} = \beta_{0jk} + \beta_{1jk}X1_{ijk} + r_{ijk},\ r_{ijk} \sim N\left(0, \sigma^2{}_{|X}\right)$$

$$\boldsymbol{L2}: \beta_{0jk} = \gamma_{00k} + \mu_{0jk}, \qquad \mu_{0jk} \sim N(0, \tau_2^2)$$

$$\beta_{1jk} = \gamma_{100}$$

$$\boldsymbol{L3}: \gamma_{00k} = \delta_{000} + \delta_{001}X2_k + \delta_{002}INT_k + \varsigma_{0k},\ \varsigma_{0k} \sim N(0, \tau_3^2)$$

where, $Y$ are the DIBELS® composite scores; $X1$ are the baseline DIBELS® composite scores; $INT$ is the treatment/control variable; $X2$ are indicator of school size (strata variable). $\mu_{0jk}$ and $\varsigma_{00k}$ are the random intercept terms for class and school respectively.

After running both models, two and three levels, we will check for any additional benefit from the three-level model (random effect for class). If the three-level model fails to converge; if the random effect variance for class is zero; or if random effects are highly correlated, we will default to the two-level model. Otherwise, we will formally test the random effect of class using a likelihood ratio test of the two nested models (two vs three level), and report the most appropriate model.

Secondary outcomes, subgroup and any additional analyses will default to the appropriate two or three-level model depending on the outcome of the sensitivity analysis for the primary outcome.

### *Secondary outcome analysis*

Components of reading fluency:

Given the distribution of most secondary outcome measures are positively scored, integers, and potentially skewed, secondary outcome analyses will use generalized linear mixed

models (GLMM; Poisson or negative binomial family depending on overdispersion) for each of the five components of reading fluency (e.g. letter naming fluency, phonemic segmentation fluency, nonsense word fluency, oral reading fluency, and word reading fluency) (10,14).

***GLMM three-level design:***

$$\boldsymbol{L1}: g(Y_{ijk}) = \beta_{0jk} + \beta_{1jk}X1_{ijk} + r_{ijk},\ r_{ijk} \sim N\left(0, \sigma^2{}_{|X}\right)$$

$$\boldsymbol{L2}: \beta_{0jk} = \gamma_{00k} + \mu_{0jk}, \qquad \mu_{0jk} \sim N(0, \tau_2^2)$$

$$\beta_{1jk} = \gamma_{100}$$

$$\boldsymbol{L3}: \gamma_{00k} = \delta_{000} + \delta_{001}X2_k + \delta_{002}INT_k + \varsigma_{0k},\ \varsigma_{0k} \sim N(0, \tau_3^2)$$

Note: $g(.) = log_e(.)$, where $g(.)$ is the log link function for the Poisson GLMM

where, $Y$ are the DIBELS® composite scores; $X1$ are the baseline DIBELS® composite scores; $INT$ is the treatment/control variable; $X2$ are indicator of school size (strata variable). $\mu_{0k}$ is the random intercept term for school.

Models for these five subscales will examine rate ratios (exponentiating the parameter estimates from the model). For each outcome, these will include the covariate of interest specified as a dichotomous variable based on whether the pupil was assigned to the treatment or control group. In addition, covariates for the baseline scores for the respective outcomes and the strata covariate will also be included.

Reading self-concept

Similarly to the other secondary outcome measures, a generalized linear mixed model (GLMM, ordinal ) will be used to predict the secondary outcome of pupil reading self-concept (including reading for pleasure/spontaneous reading). The reading self-concept measure is a bespoke measure adapted by the evaluation team based on an existing measure (see study protocol for details). The measure contains 6-items with each item scoring either 'yes' or 'no' therefore the total score use in the analysis will range between 0 and 6. For the reading self-concept measure, odds ratios will be assessed.

### *Subgroup analyses*

Subgroup analyses will first analyse the impact of the intervention on the FSM sub-group (alone), for both primary and secondary outcomes, to ascertain the size of treatment effects for FSM pupils. A further subgroup analysis will analyse the interaction of FSM and the treatment effect on the primary and secondary outcomes; to examine whether the treatment effect is conditional on FSM eligibility. We present the three-level model including class at level 2, but this will reduce to a two-level design conditional on the sensitivity analysis on the primary outcome analysis.

$$\boldsymbol{L1}: g(Y_{ijk}) = \beta_{0jk} + \beta_{1jk}X1_{ijk} + \beta_{2jk}FSM_{ijk} + r_{ijk},\ r_{ijk} \sim N\left(0, \sigma^2{}_{|X}\right)$$

$$\boldsymbol{L2}: \beta_{0jk} = \gamma_{00k} + \mu_{0jk}, \qquad \mu_{0jk} \sim N(0, \tau_2^2)$$

$\beta_{1jk} = \gamma_{100}$

$\beta_{2jk} = \gamma_{200}$

$\boldsymbol{L3}: \gamma_{00k} = \delta_{000} + \delta_{001}X2_k + \delta_{002}INT_k + \varsigma_{0k}$, $\varsigma_{0k} \sim N(0, \tau_3^2)$

Note: $g(.) = log_e(.)$, where $g(.)$ Is the log link function for the secondary outcome measures, whereas the primary outcome, $g(Y_{ijk}) = Y_{ijk}$.

### *Additional analyses*

To assess the robustness of the findings, a number of additional pupil-level covariates will be introduced: pupils' prior reading skill, school-size, pupil age, type of primary need, English as a first language (EFL), and free school meal status (FSM). These will extend thelinear mixed model analyses from the primary outcome using the same model structure and with the inclusion of the additional covariates. In addition, the interactions between the specified covariates and intervention/control group variable will be included in the model to assess whether the treatment effect is conditional on additional covariates.

### *Longitudinal follow-up analyses*

Due to the low numbers of pupils in special schools who are typically able to take part in standardised reading tests at the end of KS2, it is not anticipated that suitable data will be available for any longitudinal follow-up of our analyses. Therefore, no follow-up analyses are planned

### *Imbalance at baseline*

Characteristics of each trial arm group will be summarised descriptively, both as randomised and as analysed in the primary analysis. However, no formal statistical comparisons will be undertaken (15). Continuous measures will be reported as a mean, standard deviation (SD), minimum and maximum, while categorical data will be reported as a count and percentage.

### *Missing data*

With respect to missing data, if over 5% of cases are missing, a generalised linear mixed model will be run (16), specifying missingness as the binary outcome variable. It is plausible that data may be missing at random based on: type of primary need, EFL, and baseline pupil reading ability; and therefore, these variables will be used as covariates in the prediction of missingness. Should data be found to be missing at random, then a multiple imputation model will be run including those covariates found to be predictive of missingness. A further sensitivity analysis will compare the imputed model to the ITT analysis (14).

### *Compliance*

To ascertain the influence of non-compliance on the predictions made by the ITT, an instrumental variable approach will be utilised. Specifying two binary compliance indicators, full compliance, and partial compliance, defined as exceeding a proportion of the six binary response questions for compliance. A separate instrumental variables analysis will be

conducted for each derived compliance measure to investigate the influence of full compliance and partial compliance (Table 3).
Compliance binary response questions are defined as:

1. All Teachers/TA's supporting pupils through HERiSS have been through the digital training (this item will be scored dichotomously 'Yes/No')
2. At least one member of staff attending the two webinars (this item will be scored dichotomously 'Yes/No')
3. At least 2 trained members of staff overseeing implementation to the end of the academic year (this item will be scored dichotomously 'Yes/No')
4. Schools engaging with ISO provided fortnightly supervision (attending at least 70% of scheduled sessions) and with monthly ISO visits in schools (attending 70% scheduled sessions) (this item will be dichotomously scored, 'Yes' if 70% of supervision were held and 70% of ISO visits in schools were attended, 'No' otherwise)
5. Teachers following recommendations from ISO sessions including using the activities outlined in the HER® manual (measured using a 4-point rating completed by ISOs following each visit). (this item will be turned into a dichotomous measure with a 4 (used most of the time) as 'Yes' and 1-3 (no evidence, rarely used, sometime used) being 'No'.
6. Time- tabling 3 HER® sessions per pupil per week[4] (this item will be scored dichotomously scored 'Yes'/'No'.

**Table 3: Compliance definitions for compliance analyses**

<table>
<tr><th>Compliance items</th><th>Full</th><th>Partial (a)</th><th>Partial (b)</th></tr>
<tr><td>1</td><td rowspan="6">6 out of 6</td><td rowspan="3">2 out of 3</td><td rowspan="5">4 out of 5</td></tr>
<tr><td>2</td></tr>
<tr><td>3</td></tr>
<tr><td>4</td><td>1 out of 2</td></tr>
<tr><td>5</td><td rowspan="2"></td></tr>
<tr><td>6</td><td>✓</td><td>×</td></tr>
</table>

As recommended by the EEF, a Two Stage Least Square approach will be used to estimate the model and Huber-White standard errors reported which are robust to clustering (16). The R packages 'ivpack' and 'ivreg' will be used to implement the two-stage instrumental variable analysis (17,18). Compliance will be instrumented by the intervention allocation (19). The stage 1 model is defined as follows:

$$Compliance_k = \beta_0 + \beta_1 Treatment_k + \varepsilon_{jk}$$

Predicted values for, $Compliance_k$, from the stage 1 model will be included in the stage 2 model, as follows:

[4] The HER® team recommends 3 sessions per week

$$Y_{ik} = \beta_0 + \beta_1 \widehat{compliance}_k + \beta_2 baseline_{ik} + \beta_3 schoolsize_k + r_{ik}$$

### *Intra-cluster correlations (ICCs)*

Initially ICCs (at both the classroom and school level), will be calculated for the null model (without covariates predicting the DIBELS® composite score); and then for the primary model (i.e. the model including the baseline DIBELS® composite score, with baseline DIBELS® composite score and school size strata as covariates).

### *Effect size calculation*

Effect size for the primary outcome measure, DIBELS® composite score will be reported as Hedges' g (adjusted mean difference) (20,21). According to our three-level LMM for primary outcome, a sample estimate of the effect size equivalent to Hedges' g with 95% confidence interval (16), is defined as:

$$\hat{\Delta}_g = \frac{\widehat{\beta_1}}{S_{WT}} \sqrt{1 - \frac{2(p_u - 1)\rho_s + 2(n_u - 1)\rho_c}{N - 2}}$$

Where $\widehat{\beta_1}$ is the adjusted mean difference in DIBELS® composite score between trial arms; $S_{WT}$ is the within group pooled standard deviation (unconditional sample variance; 21)

$S^2_{WT} = \frac{\sum_{i=1}^{m^I} \sum_{j=1}^{p_i^I} \sum_{k=1}^{n_{ij}^I} (Y_{ijk}^I - Y_{i..}^I)^2 + \sum_{i=1}^{m^C} \sum_{j=1}^{p_i^C} \sum_{k=1}^{n_{ij}^C} (Y_{ijk}^C - Y_{i..}^C)^2}{N-2}$ Where '$m^I$' is the total number of schools in the intervention sample, '$p^I$' the total number of classes and '$n^I$' the total number of pupils (equivalent definitions apply for the control group, but with the 'C' designation). $Y_{i..}^I$ and $Y_{i..}^C$ are the mean outcomes among intervention and control schools respectively.

The remaining part of the $\hat{\Delta}_g$ equation makes the adjustment for clustering. The two intra - class correlation coefficients at the school ($\rho_s$) and class ($\rho_C$) level are defined as follows,

$$\rho_s = \frac{\sigma_{BS}^2}{\sigma_{BS}^2 + \sigma_{BC}^2 + \sigma_{WC}^2} = \frac{\sigma_{BS}^2}{\sigma_{WT}^2}, \qquad \rho_C = \frac{\sigma_{BC}^2}{\sigma_{BS}^2 + \sigma_{BC}^2 + \sigma_{WC}^2} = \frac{\sigma_{BC}^2}{\sigma_{WT}^2},$$

where $\sigma_{BC}^2$ is the between-class variance, $\sigma_{BS}^2$ is the between-school variance, and $\sigma_{WC}^2$ is the within-school variance. In addition, we define $p_u$ and $n_u$ as follows:

$$p_u = \frac{N^C \sum_{i=1}^{m^I} \left( \sum_{j=1}^{p_i^I} n_{ij}^I \right)^2}{NN^I} + \frac{N^I \sum_{i=1}^{m^C} \left( \sum_{j=1}^{p_i^C} n_{ij}^C \right)^2}{NN^C} n_u = \frac{N^C \sum_{i=1}^{m^I} \sum_{j=1}^{p_i^I} \left( n_{ij}^I \right)^2}{NN^I} + \frac{N^I \sum_{i=1}^{m^C} \sum_{j=1}^{p_i^C} \left( n_{ij}^C \right)^2}{NN^C}$$

Further details can be found in Hedges (2011)(21).

For the remaining secondary outcomes, their effect sizes will be reported as odds ratio (pupil reading self-concept) or rate ratios (all other secondary outcomes, exponentiated parameter estimates), given that generalized linear mixed effects models with log link function are used to model the data and that the measures are positively scored integers with some amount of skew anticipated (10,22).

# References


1. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL, https://www.R-project.org/.
2. Dong, N., & 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), 24-67.
3. Bulus, M., Dong, N., Kelcey, B., & Spybrook, J. (2021). PowerUpR: Power Analysis Tools for Multilevel Randomized Experiments. R package version 1.1.0., https://CRAN.R-project.org/package=PowerUpR
4. Allen, R., Jerrim, J., Parameshwaran, M., & Thompson, D. (2018). Properties of commercial tests in the EEF database. *EEF Research Paper*,
5. Demack, S. (2019). Properties of commercial tests in the EEF database. EEF Research Paper.
6. Grindle, C.F., Murray, C., Hastings, R.P., Bailey, T., Forster, H., Taj, S., Paris, A., Lovell, M., Jackson Brown, F. and Hughes, J.C. (2021), Headsprout® Early Reading for children with severe intellectual disabilities: a single blind randomised controlled trial. J Res Spec Educ Needs, 21: 334-344.Moerbeek, M., & Schie, S. V. (2019). What are the statistical implications of treatment non-compliance in cluster randomized trials: A simulation study. *Statistics in medicine*, *38*(26), 5071–5084. https://doi.org/10.1002/sim.8351
7. DfE, 2018: https://www.gov.uk/government/statistics/special-educational-needs-in-england-january-2018
8. Morena, L. (2011). Evaluating the impact of Headsprout on the reading achievement of English language learners (Doctoral dissertation, University of Georgia).
9. Storey, C., McDowell, C., & Leslie, J. C. (2019). Headsprout Early Reading for Specific Literacy Difficulty: A Comparison Study. Journal of Behavioral Education, 1-15.
10. Barnett, A.G., & Dobson, A.J. (2008). An Introduction to Generalized Linear Models (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315182780
11. Douglas Bates, Martin Maechler, Ben Bolker, Steve Walker (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1),1-48. doi:10.18637/jss.v067.i01.
12. Christensen, R. H. B. (2019). ordinal - Regression Models for Ordinal Data. R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal.
13. Kuznetsova A, Brockhoff PB, Christensen RHB (2017). lmerTest Package: Tests in Linear Mixed Effects Models, Journal of Statistical Software, 82(13),1-26. doi: 10.18637/jss.v082.i13 (URL: https://doi.org/10.18637/jss.v082.i13).
14. Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Analytical Methods for Social Research). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511790942
15. Senn, S. (1994), Testing for baseline balance in clinical trials. Statist. Med., 13: 1715-1726. https://doi.org/10.1002/sim.4780131703

16. EEF (2018), Statistical analysis guidance for EEF evaluations, https://d2tic4wvo1iusb.cloudfront.net/documents/evaluation/evaluation-design/EEF_statistical_analysis_guidance_2018.pdf?v=1630566986
17. Yang Jiang and Dylan Small (2014). ivpack: Instrumental Variable Estimation.. R package version 1.2. https://CRAN.R-project.org/package=ivpack
18. John Fox, Christian Kleiber and Achim Zeileis (2021). ivreg: Instrumental-Variables Regression by '2SLS', '2SM', or '2SMM', with Diagnostics. R package version 0.6-1. https://CRAN.R-project.org/package=ivreg
19. Joshua D. Angrist & Guido W. Imbens (1995) Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity, Journal of the American Statistical Association, 90:430, 431-442, 10.1080/01621459.1995.10476535
20. Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. Journal of Educational and Behavioral Statistics 32, 4:. 341 - 370 https://doi.org/10.3102/1076998606298043
21. Hedges, L. V. (2011). Effect Sizes in Three-Level Cluster-Randomized Experiments. Journal of Educational and Behavioral Statistics, 36(3), 346–380. http://www.jstor.org/stable/29789487
22. Hayes, R.J., & Moulton, L.H. (2017). Cluster Randomised Trials (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.4324/9781315370286