

# Statistical Analysis Plan

TRIAL FULL TITLE	Evaluation of 'Graphogame Rime'
DEVELOPER	University of Cambridge and University of Jyväskylä Finland
EVALUATOR	National Foundation for Educational Research
TRIAL REGISTRATION NUMBER	ISRCTN10467450
TRIAL STATISTICIAN	Jack Worth
TRIAL CHIEF INVESTIGATOR	Ben Styles
SAP AUTHORS	Jack Worth and Ben Styles
SAP VERSION	3
SAP VERSION DATE	13 <sup>th</sup> September 2016
EEF DATE OF APPROVAL	15 <sup>th</sup> September 2016
DEVELOPER DATE OF APPROVAL	1 <sup>st</sup> July 2016

## Introduction

The aim of the National Foundation for Educational Research (NFER)'s evaluation of Graphogame Rime is to test the impact of the programme on pupils' reading skills. Graphogame Rime is a computer game developed to teach pupils to read by developing their phonological awareness and reinforcing the connections between phonological units and spelling units, highlighting rhyming connections. Originally developed by a Finnish University, the Graphogame group of computer programs employ algorithms that analyse a child's performance and constantly adjust the difficulty of the content so that the challenge matches the learner's ability.

The primary research question is whether playing Graphogame Rime as part of literacy lessons improves pupils' development of reading ability, compared to 'business as usual' participation in literacy lessons.

The NFER/Cambridge research team aims to answer this question by recruiting ten primary or infant schools to participate in the study and randomly allocating pupils each to one of two groups: a group of pupils that spend 10-15 minutes each day playing on the Graphogame

Rime computer games during literacy lessons and a 'business as usual' control group of pupils from the same classes who will have normal tuition in literacy lessons.

The University of Cambridge ('Cambridge') research team administered tests in literacy (New Group Reading Test (NGRT) – Level 1B) to selected year 2 pupils in all schools to measure literacy ability at baseline. The NFER research team has administered post-tests in literacy (NGRT – Level 1B) to the same pupils to measure the development of literacy ability.<sup>1</sup> The NFER research team have also administered a spelling test (Single Word Spelling Test) to measure the development of spelling ability.

The aim of the analysis is to measure the differential progress in literacy development between pupils in each group, and thereby measure the effect of the training programme and teaching materials when administered by teachers under typical classroom conditions.

## **Sample size**

The aim of the study was to recruit 10 Cambridgeshire schools and include year 2 pupils that had not reached the expected standard in the phonics screening check. Eligible pupils would then be randomly allocated to either the intervention or control group (see 'Randomisation'). The same design will be used for two years (2015/16 and 2016/17), doubling the overall sample size.

Initial estimates of statistical power suggested that a total sample size after two years of 400 pupils would be sufficient to detect a standardised effect size of 0.17 with 80% power.<sup>2</sup> For sample size calculations, we assumed a final model without a school-treatment interaction and including a fixed effect for school. This is appropriate since we do not intend to generalise the results of the trial beyond its schools (Feaster et al, 2011). This effect size is smaller than might be expected of the Graphogame programme according to previous small-scale researcher-administered studies, but is appropriately conservative given that previous research had used outcome measures closely aligned to the intervention and that individual researchers sat with children and encouraged them to persist on task while they played Graphogame.

Cambridge was able to recruit 11 schools in the first year, from which 216 pupils were tested at baseline and allocated to a group. Given that we might expect up to 10 per cent attrition, this is in line with our expectations of the achieved sample size for the first year. We anticipate the sample size being similar in the second year of the study.

## **Randomisation**

An NFER statistician performed a stratified randomisation of pupils by class with the aim of balancing the randomisation at class level (i.e. having equal numbers in intervention and control groups within each class) to aid delivery. We first put schools in a random order; within school we put classes in a random order; and within classes we put pupils in a

---

<sup>1</sup> Use of the Level 1B test was a change to the original protocol. See 'primary outcome' section.

<sup>2</sup> Assuming a two-tailed test with a 95% confidence level, number of schools=10, expected number of pupils per school=20, intra-school correlation=0 (pupil randomisation), correlation between pre- and post-test scores=0.8.

random order. The number of classes per school varied from one to three. The entire list was then allocated to control and intervention in turn: first pupil to control, second to intervention, third to control, and so on. The SPSS syntax used to randomise pupils in the first year is shown in Appendix 1, and similar syntax will be used in the second year.

## **Outcome measures**

### **Primary outcome**

The primary outcome in the trial is pupils' test scores on the New Group Reading Test (NGRT) Level 1B. The original protocol stated that the primary outcome measure would be NGRT Level 2. A protocol amendment was made in March 2016 following the pre-test because we expected pupils to find the content in NGRT Level 2 too difficult, given how difficult they had found the pre-test.

The post-tests were (and will be in the second year) administered by NFER test administrators to ensure independence from the Cambridge research team that are delivering the intervention. Test administrators were not informed of which group each pupil has been allocated to and were advised not to discuss the interventions with teachers and to conduct the administration as they normally would. Post-testing for the first year of the evaluation took place during 11<sup>th</sup> – 25<sup>th</sup> April 2016.

We will use raw test scores (i.e. number of correct answers) rather than standardised or age-standardised scores for the analysis. The expected range of the scores is 0-43. We expect within-year age to be evenly distributed across the two groups since they have been randomised; processing the scores further would be of limited benefit for this research and might introduce ceiling or floor effects.

### **Secondary outcomes**

A secondary outcome identified in the protocol will be used to test the secondary hypothesis that Graphogame Rime leads to an improvement in spelling ability. NFER test administrators will administer the GL Assessment Single Word Spelling Test A at the same time as NGRT and we will analyse the raw score.

## **Analysis**

### **Overview and definitions**

This evaluation is an efficacy trial conducted within 11 schools in Cambridgeshire, many of whom are familiar with working with the developer. Schools were selected non-randomly and there is no intention of attempting to generalise the results of the trial beyond the participating schools. We are therefore in a situation of 'conditional inference' (Hedges et al, 2009); inferences are restricted to the particular schools in the study. This has considerable implications for the choice of analysis model. We shall model class<sup>3</sup> as a fixed effect. Additionally, we will model the interaction between class and intervention as a fixed effect.

---

<sup>3</sup> Randomisation was restricted by class rather than school – see 'randomisation' section.

There is considerable debate in clinical trials about how to report the main intervention effect when there is a significant site by treatment interaction (Feaster et al, 2011). For the purposes of this study, the main effect will be from a model without interaction terms i.e. the weighted mean of treatment effects across sites.

### **Standardised effect size**

The coefficients of the statistical model are measured in terms of raw test score. As is standard for EEF evaluations, the effect size will be calculated as the coefficient on the intervention group indicator divided by the total sample standard deviation. We will apply Hedges' adjustment for small sample bias of estimated variance.

### **Confidence intervals**

We will estimate a 97% confidence interval (see section on interim reporting, below) alongside the standardised effect size to give the precision with which the effect size has been estimated. The upper and lower bounds of the confidence interval will be calculated as the effect size plus/ minus the product of the critical value of the t-distribution and the standard error of the intervention group indicator coefficient estimated from the model.

### **Primary analysis**

The primary analysis of the impact of the intervention will be a regression model of reading test score at post-test including all data from both cohorts. The model will be estimated with all pupils that completed a pre-test and a post-test.

The primary outcome model will have raw score at post-test as the dependent variable and the following covariates will be included in every model:

- an indicator of whether the pupil is in the Graphogame intervention group. The excluded group is the control group, so the coefficient of the intervention group indicator measures the difference in (conditional) outcomes between the intervention group and the control group
- the pupil's raw score on the pre-test. The coefficient is incidental to the research, but explains a large proportion of outcome variance, increasing the power of the analysis.
- an indicator of which class and intervention cohort the pupil was in, to account for the stratified randomisation. As classes are nested within schools, an indicator for school is not necessary for this single-level model.

The above model will yield the main intervention effect for publication in the executive summary. In addition, a further model will explore the interaction between class and intervention in order to determine whether Graphogame was differentially effective between different classes or schools. This analysis will be somewhat underpowered but will yield interesting results in terms of whether certain classes or schools had larger effects than others.

The sample size and the rate of school and pupil attrition (i.e. the number of schools and pupils analysed compared to the number of schools and pupils that were randomised) will be reported in the CONSORT flow diagram. In an effort to determine whether there was any biased attrition, a summary of the analysed pupils' average characteristics will be reported by randomised group.

## Interim reporting

An interim analysis will take place at the end of the first year on just the primary outcome. The interim results will be discussed with the EEF Grants Committee, with advice from the Evaluation Advisory Group if required, to determine whether or not the trial should continue. If there is substantial evidence of a positive effect of the intervention then the Grants Committee may advise to move to a larger effectiveness trial, while if there is evidence of negative effects then they may advise to stop the trial. However, the analysis of data from the first year will be underpowered for plausible positive or negative effects and therefore it is most likely that the trial will be completed as planned.

We will use the following rule to decide what to recommend on the basis of the interim analysis:

Interim result	Decision
Significant positive effect ( $p < 0.0294$ )	Expand to effectiveness trial
No significant effect ( $p > 0.0294$ )	Continue as planned
Significant negative effect ( $p < 0.0294$ )	Stop trial

Armitage et al. (1969) were the first to numerically compute the extent to which the type I error probability is increased over its nominal level if a standard hypothesis test is conducted at each of a series of interim analyses. They studied the problem of testing a normal mean with known variance and set the significance level (or type I error probability) for the trial to be 5%. If one interim analysis and one final analysis are performed this error rises to 8%. This is the situation here. Although the final result will not be predicated on a significance level, the decision at the interim analysis is binary and we should therefore adopt an appropriate threshold to aid our decision. The threshold  $p < 0.0294$  is the Pocock boundary for two planned analyses (Pocock, 1977). Were the final analysis to use a significance threshold, it would also need to be  $p < 0.0294$ . We will hence construct 97% confidence intervals around the final result and note that any reference to the trial demonstrating a 'significant' result (at either analysis) is still at the conventional level of  $p < 0.05$ .

## Missing data

Missing data presents a problem for analysis, whether a pupil is missing a value for an outcome variable (post-test score) or for covariates (e.g. pre-test score). If outcome data is 'missing at random' given a set of covariates then the analysis has reduced power to detect an effect; if data is 'missing not at random' (for example, differential dropout in the intervention and control groups for unobserved reasons) then omitting these pupils (as in the

primary ‘completers’ analysis) could bias the results. Imputing missing data could improve the robustness of the analysis and examine how sensitive the results are to alternative assumptions.

If the amount of missing data exceeds 5% of the randomised sample then we will consider using multilevel multiple imputation methods to check whether the results from the ‘completers’ analysis are sensitive to missing data. A discussion of the results in the context of missing follow-up data will be presented. If follow-up data is missing at random given covariates, and these covariates are included in the model, the results will be unbiased. It may be that the results of the multiple imputation do not differ appreciatively from the completers analysis. If this is the case and we are reasonably confident that covariates explain any missingness then this will complete the primary analysis. Otherwise, some sensitivity analysis (e.g. using extreme values) may be necessary.

### **Pre-specified secondary analysis**

**1. Analysis of differential impact on FSM pupils:** the primary analysis models will be estimated for the sub-sample of pupils that have been FSM since starting school (the indicator of Pupil Premium status). The effect size will be measured in the same way as above. The standard deviation will be re-estimated for the FSM-only sample to use for standardising the estimated effect size. We will also estimate an FSM-by-intervention interaction model to analyse the differential impact of the intervention on FSM pupils.

**2. Analysis of differential impact by pupil ability:** the primary analysis model will be run with the addition of an interaction between pre-test score and the intervention group indicator.

**3. Impact on spelling:** the primary analysis models will be run with Single Word Spelling Test raw scores as an outcome variable.

### **Sensitivity Analysis**

We will carry out analysis of the primary outcome using a multilevel model with class and class-by-intervention interaction as random effects. Class will be uniquely numbered. We will test whether the class by intervention interaction is significant. If not, the effect size itself (without the interaction term) should be the same as that from the fixed effects model. If it is significant, this will tell us something about the extent to which the effectiveness of the intervention varies depending on which class it is delivered in. Were we trying to generalise to a wider population of schools, this would be important to take into account. Since we are not seeking to generalise results beyond the schools in this effectiveness trial, the effect size and confidence interval from the fixed effects model (without interaction) will still serve as the definitive result of the trial.

### **Exploratory analysis**

Exploratory analyses will focus on the implementation of the intervention to attempt to understand *how* the intervention worked rather than *whether* it worked. Differences in the way the intervention was implemented across schools were at the discretion of school staff and researchers on the Cambridge team, so this analysis cannot be definitively interpreted

as causal. However, the analysis may give clues as to what drives any effect that may be identified by the impact analysis. It will investigate whether different approaches to implementation led to different amounts of progress made by pupils.

We will conduct an on-treatment analysis of the number of hours spent using Graphogame. Pupil-level game usage data will be recoded into levels of programme fidelity (shown in Table 1). We will explore the differential impact of the intervention by fidelity level by including indicator variables for each fidelity level as well as the intervention group indicator. It will also be analysed without recoding, to give an indication of the amount of impact an additional hour of using Graphogame made to reading.

Whilst this analysis appears attractive, it is very vulnerable to bias as those individuals who used the program the most are likely to have other characteristics that are associated with improved test performance. To mitigate this, Complier Average Causal Effect analysis (Sussman and Hayward, 2010) will be carried out using a continuous variable to describe dosage.

Table 1 Classification of programme fidelity

Level of programme fidelity	Total number of hours spent on Graphogame Rime
Low	0-3
Medium	3-6
High	6-10
Very high	More than 10

## References

Armitage P, McPherson CK, Rowe BC. (1969) *Repeated significance tests on accumulating data*. J Roy Statist Soc Series A 1969; 132: 235±244.

Feaster, DJ; Mikulich-Gilbertson, S and Brincks, AM (2011) *Modeling site effects in the design and analysis of multisite trials*. Am J Drug Alcohol Abuse. 37(5): 383–391.

Hedges, L and Rhoads, C (2009). *Statistical Power Analysis in Education Research* (NCSE 2010-3006). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. This report is available on the IES website at <http://ies.ed.gov/ncser/>.

Pocock SJ (1977). *Group sequential methods in the design and analysis of clinical trials*. Biometrika 64 (2): 191–9.

Sussman JB and Hayward RA (2010). *An IV for the RCT: using instrumental variables to adjust for treatment contamination in randomised controlled trials*. BMJ 2010; 340:c2073

## Appendix 1 – Randomisation syntax

Title 'Pupil randomisation'.

get file = "I:/EERR/First year/Definitive list of eligible pupils.sav".

\* Set randomisation seed.

set rng=mt, mtindex=08122015.

\* Randomise school order.

compute schrand=rv.uniform(0,1).

sort cases by NFERNO.

aggregate outfile = "c:/temp/schools.sav" /break = NFERNO

    /sch = first(schrnd).

match files /file = \* /table = "c:/temp/schools.sav" /by NFERNO.

\* Randomise class order.

compute clsrand=rv.uniform(0,1).

sort cases by NFERNO Class.

aggregate outfile = "c:/temp/class.sav" /break = NFERNO Class

    /cls = first(clsrnd).

match files /file = \* /table = "c:/temp/class.sav" /by NFERNO Class.

\* Create variable for randomising pupils.

compute rand=rv.uniform(0,1).

\* Stratify randomisation by class.

sort cases by sch cls rand.

\* Allocate pupils to groups by order.

compute group = MOD(\$casenum,2).

variable labels group = "Group allocation".

value labels group 0 'Control group' 1 'Intervention group'.

\* Group split overall.

frequencies group.

\* Group split by school/class.

crosstabs NFERNO by group.

crosstabs Class by group by NFERNO.

sort cases by NFER\_PUPIL\_ID.

save outfile = "I:/EERR/First year/Randomisation.sav".