Statistical Analysis Plan for

"Digital feedback in Primary Maths"

RAND Europe

PROJECT TITLE	**Digital feedback in Primary Maths**
DEVELOPER (INSTITUTION)	James Siddle (Kyra School Alliance)
EVALUATOR (INSTITUTION)	RAND Europe
PRINCIPAL INVESTIGATOR(S)	Julie Belanger
TRIAL (CHIEF) STATISTICIAN	Alex Sutherland
SAP AUTHOR(S)	Alex Sutherland & Janna van Belle
TRIAL REGISTRATION NUMBER	ISRCTN50447665
EVALUATION PROTOCOL URL OR HYPERLINK	*https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/Round_10_-_Kyra.pdf*

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]	06/06/2018	

Table of Contents

Introduction 3
Study design 4
Randomisation 5
Power calculation and sample size 5
Outcome measures 7
    Primary outcome 7
    Secondary outcome 8
Analysis 8
    Primary intention-to-treat (ITT) analysis 8
    Interim analyses 9
    Imbalance at baseline for analysed groups 9
    Secondary outcome analyses 10
    Effect size calculation ..... **Error! Bookmark not defined.**
    Missing data 11
    Non-compliance with intervention 11
Report tables 12

Introduction

Description of the intervention

The aim of this trial is to increase the specificity and relevance of teachers' feedback using a digital app and tablet, making it easier for pupils to review and respond to feedback. The use of new technology is innovative but the underlying model is one of improving the feedback that teachers give their pupils (aka providing effective feedback). The approach was originally developed and tested by James Siddle for the Kyra Teaching School Alliance (Siddle 2015)).

There is a single intervention package being randomised in this trial. The package consists of: (I) training for teachers to improve feedback; (II) training on the use of tablets for providing this feedback digitally; (III) support from school 'Research Leads'; (IV) regular meetings to share and discuss practices (Professional Learning Communities (PLC) meetings). The primary focus of the trial is the overall effect of this package on pupil attainment in maths. The control condition will be 'business as usual'. This statistical analysis plan sets out how we will assess whether this approach ('Digital Feedback') leads to improvements on pupil maths outcomes compared to 'business as usual'.

As part of the intervention, teachers will conduct two main activities. First, teachers will receive diagnostic assessment resources to quickly assess their pupils' understanding of a topic, for example, through multiple choice questions. These assessments consist of only 3 or 4 questions to quickly assess the level development of a pupil. These assessments can be used as often as teachers deem useful and necessary, and will help teachers identify pupils that have a greater learning need, and tailor the delayed feedback to the needs of these pupils. The training guidelines suggest using diagnostic assessments 2 or 3 times a week. Second, teachers will provide feedback on the mathematics books that pupils are using as part of the curriculum using a tablet (hence 'digital'). Feedback will be given throughout the school year on any type of mathematics classwork. Digital feedback consists of teachers recording video summaries of feedback, rather than writing down comments. Feedback is of two types: immediate and delayed. Immediate feedback is read and used by pupils while they are conducting their work in class. Delayed feedback can be reviewed by pupils later in time and is intended to be primarily used for pupils who are struggling with a topic, or who may have difficulties with mathematics more generally. The results from the diagnostic assessment conducted under the first step can be used to tailor the delayed feedback to pupils with an apparent greater need for support.

The duration of the intervention is 13 months; from June 2017 when the teacher training sessions took place, to July 2018 when the post-intervention surveys will take place. Training on when and how to use feedback, and the use of ICT to provide feedback (part I and II of the intervention) was provided to teachers in a one-day training session which was organised on different days in June and one additional training day in September. One-day training for the Research Leads (part III of the intervention) was given on three different days throughout June 2017. Teachers will take part in monthly "professional learning communities" throughout the 2017-2018 school year (part III of the intervention), which take place monthly, with multiple different dates available each month to accommodate teachers. The Research Lead in each school will provide ongoing support and coaching to the teachers (part IV of the intervention). Coaching and general support is included as part of this trial, to help teachers effectively use the app and improve the quality of their feedback.

Design overview

Trial type and number of arms		Two-arm, cluster randomised
Unit of randomisation		Class
Stratification variables (if applicable)		-
Primary outcome	variable	Maths Attainment
	measure (instrument, scale)	ACERs Essential Learning Metric (ELM) for Mathematics for Years 4 and 5
Secondary outcome(s)	variable(s)	Pupil Engagement
	measure(s) (instrument, scale)	Engagement versus Disaffection survey

Study design

The study is a two-arm cluster-randomised control trial (RCT) in 34 primary schools in Lincolnshire, North Lincolnshire, Nottinghamshire, and the Humber. Schools were recruited by the Kyra School Alliance based on their capacity to recruit within these areas but not beyond. Based on assumptions about school/class size, there should be 2,700 pupils involved, distributed over 34 schools and 108 classes.

The following eligibility criteria for participating schools apply:

- Only state schools are included.
- Schools should have enough tablets for one year group/class (roughly 1 tablet for every 4 or 5 pupils).
- Schools should not have had the Digital Feedback intervention before.

Our initial plan was to run a cluster-randomised trial, with schools as the unit of randomisation, but recruitment and delivery limitations meant this was not feasible. The compromise was to implement a within-school design, randomising classes.[1] In schools where teachers teach a combined Year 4 and Year 5 class, this combined class will be included as a single year group in the randomisation and analysis.

There is only one treatment condition in this trial. Year 4 and 5 pupils in treatment classes will receive digital feedback in the school year 2017/18. For control classes, it will be business as usual during the school year 2017/18.The aim of the trial (and our analysis) is to assess whether the feedback intervention leads to an improvement in maths performance greater than that observed in business-as-usual classrooms. All classes that are part of the trial had a 50% chance of being assigned to the treatment condition.

The baseline data consist of Key Stage 1 mathematics scores, which were collected either 1 (for year 4 pupils) or 2 (for year 5 pupils) years before randomisation. The outcomes for all pupils will be assessed once at the end of the trial, in June 2018, with an external mathematics

[1] Year-group based randomisation was also discussed, but was not feasible owing to too few year-groups to power the study (discussed further in the power and sample size section).

test (ACERs Essential Learning Metric (ELM) for Mathematics) which will be independently administered and invigilated.

Randomisation

The unit of randomisation was classes in year 4 and year 5, i.e. year 4 or 5 classes in a given school will be randomly assigned to treatment or control.

Allocation to treatment or control was conducted on the 12th of June 2017 by Dr Sutherland, and included all classes for which baseline data were available and the school had signed the Memorandum of Understanding (MoU). A stratified randomised design was planned, with classes in year 4 and year 5 as the unit of randomisation. However, misunderstandings between schools and Kyra during recruitment, only disclosed to the evaluation team until just prior to randomisation, meant that the research team had to ensure each school was allocated at least one treatment class. That in turn meant simple stratified randomisation was not possible. To accommodate this request, and to ensure each school will have at least one treatment class, randomisation took place separately within each school. Due to the higher than anticipated number of small schools (schools with only one or two year groups), stratification within schools was considered undesirable.

In practical terms, we assigned random numbers to the list of classes, sorted the data by school, and assigned the first (randomly ordered) class within each school the treatment condition and the second to the control. Subsequent classes were then assigned to treatment and control sequentially, with these steps undertaken using code run in Stata so that although 'sequential', allocations were instantly completed. This approach assumed schools had equal number of year 4 and year 5 classes, to avoid unbalanced treatment allocation. In reality there were more schools with odd numbers of classes, which resulted in slightly more treatment allocations than control allocations. Table 1 below shows the allocations by class year group:

Table 1: class allocations by year group of class

Year	Control	Treatment	Total
4	18	28	46
5	26	21	47
3&4 combined	3	3	6
4&5 combined	0	1	1
5&6 combined	4	4	8
Total	51	57	108

Power calculation and sample size

The trial is intended to focus on analysis for all-pupils as the main effects analysis. As an exploratory analysis we would like do sub-group analyses of year 4 and 5 pupils if the numbers allow, however we acknowledge that this analysis is likely to be underpowered. The aim was that the trial would consist of 120 year groups over 60 schools; with 60 year groups (30/30 year 4/5) in each treatment arm.

Using these as starting points, and using data provided by Kyra and the EEF on the average number of pupils per school in the regions identified, we assumed that there are on average 25 pupils per class. We assumed 1 class per year group, and two year groups (4 and 5) per

school. We further assumed, based on EEF guidelines[2] and requirements for power calculations, that the amount of variation explained by covariates is 0.53 for level 1 and 0.00 for level 2, and that the intra-cluster correlation coefficient (ICC) is 0.13 (13%). Lastly we also assume an alpha of 5% and an intended 80% power to detect effects. Power and minimum detectable effect size (MDES) calculations were performed using PowerUp! (Dong and Maynard, 2013). Table 2 sets out the MDES calculations.

Based on the assumptions set out above, the initial, between-year based MDES calculation was 0.197 (Column A of Table 2). However, Kyra eventually recruited 34 schools (or 67 year groups – one school had a split year 4/5 class, which will be treated as a year 5 class in the analysis). Because the school and year group recruitment was below target, it was decided (after discussion between RAND Europe, Kyra and the EEF in May 2017) to change the randomisation plan from *year group* randomisation to *class level* randomisation. Because of the presence of a few large schools in the sample, this meant that the same number of schools now provided 108 classes for randomisation.

Given that we were randomising within schools, the ICC represents variation between *classes* rather than variation between *schools*, but we had no national data on which to base the between-class ICC. Instead, we used two assumed ICCs with the actual sample. For our calculation with the actual sample of classes (Column B in Table 2) we assumed the ICC would be 10%. For the second calculation (Column C in Table 2), we assumed this would be 20%. Table 2 shows that on the basis of the above assumptions, n=108 classes would provide an MDES of 0.186 for an ICC of 10% and an MDES of 0.252 for an ICC of 20%.

Table 2: MDES calculation for each year group in the trial

	A	B	C
UoR*	**Initial calculation Year groups**	**Actual sample Classes**	**Actual sample Classes**
Year groups	120*	67	67
Average number of classes within schools across year 4 and 5	2	3.2	3.2
Unit of randomisation: Classes (total)	120	108+	108+
Pupils per class (assumed)	50*	25+	25+
Intra-class correlation [ICC] (assumed)	0.13*	0.10*	0.20+
L2 covariates	N/A	N/A	N/A
Variance explained by L1 covariates (assumed)	0.53*	0.53*	0.53*
MDES	0.197	0.186	0.252

Table notes: Column A= initial sample calculation; Column B achieved sample size with an ICC of 0.10;. * Items used for initial MDES calculations. + Items used for MDES calculations with sample of classes used for randomisation.

Allocations and withdrawals to date

Classes were randomised (using procedure outlined above) to either the control or treatment groups. Following random allocation, there were 57 classes in the treatment group, and 51 classes in the control group.

[2] https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Protocol/ICC_2015.pdf

Randomising within year groups means an increased risk of contamination (controls receiving intervention) as a result of having intervention and control classes in the same year. This risk is limited with regards to using the feedback app, as that will only be provided to treatment teachers and not be available to control teachers. However, contamination is more of a concern with regards to the use of diagnostic assessments by teachers (which are publicly available), and the coaching that teachers receive. To minimise this risk, teachers and Research Leads will be provided with guidelines that make explicit this requirement and underline the importance of separation between treatment and control classes. Adherence to this requirement will be investigated as part of the process evaluation. Further, the two surveys at midline and endline will seek to understand the feedback practices of both treatment and control teachers so as to compare changes over the course of the trial.

At the time of writing (August 2017) 10 of the 108 intervention classes had withdrawn from the trial. The schools, together with the reasons cited by them, and the number of affected classes is listed in table 3. The list of reasons cited for withdrawal will be fully reported following CONSORT standards in the final report, using monthly update information provided by Kyra. One observation about rows one and five, 'number of intervention classes higher than expected by school', is that this suggests schools did not fully understand what they were signing up for.

Table 3: reasons for withdrawal given by schools (to date)

#	Reason for withdrawal	Number of classes (intervention /control)
1	Number of intervention classes higher than expected by school	2 (2/0)
2	RL has too little capacity	3 (2/1)
3	Unknown	2 (1/1)
4	Unknown	2 (1/1)
5	Number of intervention classes higher than expected by school	6 (3/3)
6	RL has too little capacity	2 (1/1)
		17 (10/7)

.

Outcome measures

Primary outcome

The primary outcome measure at the end of Years 4 and Year 5 will be a measure of mathematics, administered and collected by the Australian Council for Educational Research (ACER). The study will use ACERs Essential Learning Metric (ELM) for Mathematics for Years 4 and 5. The tests will be paper-based and will take 45-50 minutes to complete.

The tests consist of two sets of measures: mathematical content (number, algebra, measurement, geometry, statistics and probability); and mathematical processes (understanding, fluency, problem solving and reasoning).The tests have been reviewed by the Kyra School Alliance for compatibility with the mathematics curriculum taught and it was agreed that these tests would capture the key elements that feedback would focus on. The tests will be independently delivered, invigilated and marked by ACER, with data then supplied to RAND Europe for analysis. Schools will not be told in advance what the test is (i.e. the name

of the test) but will be informed about the general areas it covers. Test invigilators will be blind to treatment allocation.

During analysis, the treatment/control coefficient will relate to the difference between treatment and control classrooms in raw scores derived from the ACER ELM assessments.[3] We expect the coefficient to be positive, meaning that treatment classes have higher scores on ELMS than control classes.

Secondary outcome

As a secondary outcome we will conduct a pupil engagement test which will be given to all pupils at the end of the trial together with the mathematics outcome test. Pupil engagement is one of the key mechanisms through which digital feedback is hypothesised to improve pupil outcomes, and measuring pupil engagement is therefore crucial to understanding the efficacy of the intervention. The instrument to capture pupil engagement we will use is the Engagement versus Disaffection survey developed by Skinner et al (2009). It captures behavioural and emotional engagement and dissatisfaction in the classroom and has a high score reliability (0.79), adequate cross-time correlation (0.63) and high factor loadings on the factors (ranging .55-.84). Further detail on the psychometric properties of the test are reported at length in Skinner et al (2009).

Analysis

Primary intention-to-treat (ITT) analysis

The analysis of primary outcomes (as detailed above) for both Year 4 and Year 5 pupils will be on an intention-to-treat (ITT) basis[4]. This method compares outcomes means for the treatment and comparison groups, and subjects are analysed according to their randomised group allocation regardless of whether or not they adhered to this group. The ITT approach is inherently conservative as it captures the averaged effect of *offering* the intervention, regardless of whether or not the participants comply with the assignment. Problems of dropout/non-attendance may be an issue for this trial depending on how motivated teachers are. There is also the risk that schools and/or teachers will drop out at some point during the trial, as has already occurred.[5]

The unit of analysis here would be pupil level outcomes for those in classes with and without Digital Feedback being used. We will explicitly account for clustering of pupils by class via multilevel models (i.e. random effects models). Multilevel models assume that higher level units in the study are a random sample of all units – which is sometimes a source of contention – but one benefit of this approach is being able to explicitly partition variance and more flexibly

[3] We will receive the raw scores from ACER and investigate the distribution of raw scores for the presence of floor/ceiling effects. If there are present, we would use standardized scores instead.

[4] https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Research_Report/2015_Analysis_for_EEF_evaluations.pdf

[5] One way of handling selective attrition – selective exposure to treatment – is the instrumental variables (IV) approach proposed by Angrist et al (Angrist, Imbens, & Rubin, 1996).

handle complex variation (Snijders & Bosker, 2012). We would account for school-level clustering via robust standard errors.

The general equation for the multilevel model is given below:

$$y_{ij} = \propto + X_{ij}\beta + Z_i b_j + \delta DF_j + u_{ij} + u_j \qquad i = 1..N, j = 1. \quad (1)$$

Where y_{ij} denotes the pupil level outcome; i and j denote pupil and class indexes respectively; X_{ij} is the baseline attainment data. DF_j is a dummy variable denoting treatment/control group at class level; Z_j indicates class characteristics, in this case the year group variables. Finally, u_{ij} is the pupil-level error term, and u_j is the class-level error term. In equation (1) α denotes the overall model intercept. u_j are referred to as class-level residuals ($u_j \sim i.i.d\ N(0, \sigma_u^2)$) and e_{ij} are individual-level residuals ($e_{ij} \sim i.i.d\ N(0, \sigma_e^2)$). In the event of analysis problems arising through the overlap between year group variables and the number of classes we will omit the year group variable but will report results for both sets of analysis.

The coefficient δ associated with the DF_j dummy will constitute the main result of the trial. "All pupils" represent the main analysis, so to obtain the main result of the trial, Equation (1) above will be estimated for all pupils.

The baseline pupil measure we will include in our analysis is a continuous variable for attainment, which is operationalised as Key Stage 1 prior attainment results in Maths (NPD variable will be KS1_MATPOINTS). These were collected either one or two years before randomisation (for Year 4 and Year 5 pupils respectively). These pupil data will be obtained from two sources. Initially, and prior to randomisation, from schools: lists of pupils in the target year groups have been provided by all in-trial schools before schools were made aware of how their classes were allocated to control or treatment conditions. Secondly, from linkage to NPD. All analyses will be performed in Stata, version 13.1 onwards.

Interim analyses

No interim analyses are planned.

Imbalance at baseline for analysed groups

A well-conducted randomisation will, in expectation, yield groups that are equivalent at baseline (Glennerster & Takavarasha, 2014). Because classes within schools here are randomly allocated to the control and intervention conditions, any imbalance at baseline will have occurred by chance. To assess imbalance at baseline, we will compare groups at class and pupil levels. At the class level, the comparison will look at the following variables, by means of cross-tabulations and histograms that assess the distribution of each characteristic within the control and treatment groups:

- Teacher experience (years of experience)
- Teacher's age
- Teacher's gender

At the pupil level, baseline balance will be assessed for the following characteristics:

- Age (Year group)

- Eligibility for FSM
- Gender
- KS1 attainment

Statistical significance tests will not be carried out to assess the balance, as their premise does not hold in randomised control trials[6] (i.e. given appropriate randomisation procedures were followed, any differences between control and treatment groups at baseline will be by definition due to chance, and classical statistical testing is therefore unnecessary). Instead, tables of the class-means (and standard deviation, where appropriate) for each characteristic will be presented, and the magnitude of any differences explored.[7] For skewed variables then quartile based measures will be used.

Further analyses to be undertaken for the final report:

- Analysis of the characteristics of the school and pupil sample in relation to the characteristics of all schools in England (available at aggregate level), to assess the representativeness of the study sample and potential for scale up.

Secondary outcome analyses

Pupil engagement, as measured with the Engagement versus Disaffection survey, will be compared between classes using the same model specification as above, substituting the secondary outcome, and without baseline attainment in the model, as baseline scores are not available.

Effect size calculation

With the multilevel models we will use the effect sizes for cluster-randomised trials given in the EEF evaluator guidance, an example, adapted from Hedges (2007) is given in Equation 1:

$$EQUATION\ 1: ES = \frac{(\bar{Y}_T - \bar{Y}_C)_{adjusted}}{\sigma^2}$$

Where $(\bar{Y}_T - \bar{Y}_C)_{adjusted}$ is the mean difference between intervention groups adjusted for baseline characteristics and σ^2 is the population standard deviation (variance), which we would request from the EEF and/or the Department for Education. The ES therefore represents the proportion of the population standard deviation attributable to the intervention (Hutchison and Styles, 2010). A 95% confidence interval for the ES, that takes into account the clustering of pupils in classes, will also be reported and will be calculated by dividing the upper and lower confidence interval bounds by the population standard deviation.

If the population variance is not available we would substitute the sample-based, pooled unconditional variance (s*) in the denominator, which is an estimate of the population variance. In multi-level models this variance will be the weighted average of the variance of treatment and control groups, as per EEF guidance (EEF, 2018: 4 footnote 11). The formula for the denominator is given in Equation 2:

[6] http://www.consort-statement.org/checklists/view/32-consort/510-baseline-data

[7] There is a convention in some disciplines that a 10pp (or larger) difference in treatment and control means at baseline constitutes 'imbalance' is thus justification for including those measures in sensitivity analyses.

$$EQUATION\ 2: s^* = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

If there is reason to believe that DF has affected the spread of the distribution in the treatment group (dispersion),[8] the unadjusted variance of the control group would be substituted into the denominator.

Missing data

Missing data can arise from item non-response or attrition of participants, and/or teachers and / or schools. We will first determine the proportion of missing data in the trial. Our use of administrative data for pupil baseline data should reduce missingness arising from both item non-response and attrition.

Our analysis would always begin with an intention-to-treat approach as set out above, meaning that we would look to analyse all classes and pupils as allocated at randomisation, regardless of compliance, and then move from there depending on the extent of missingness and where it occurs in the data.

We will explore attrition across trial arms as a basic step to assess bias (Higgins et al., 2011). We will provide cross-tabulations of the proportions of missing values on all baseline characteristics (as detailed in the previous section, at both pupil and school level), as well as on the primary outcome measures.

To assess whether there are systematic differences between those who drop out and those who do not – and thus whether these factors should be included in analysis – we will model missingess at follow-up as a function of baseline covariates, including treatment. The analysis model for this approach will be a logistic model that mirrors the multilevel level model given above (pupils clustered in classes), but the outcome will be missing to follow-up yes/no.

For less than 5% missingness overall a complete-case analysis might suffice (i.e. assuming data are MCAR), but our default will be to check results using approaches that account for missingness but that rely on the weaker MAR assumption. Our preference is to use Full-Information Maximum Likelihood (FIML) over multiple-imputation because FIML can be estimated in a single model and simulation studies show that it can reduce bias as well as MI (for a discussion of FIML vs MI see Allison, 2012). (For missingness on outcome variables only then standard statistical packages such as Stata use ML for estimating parameters so FIML would not be necessary (Allison, 2012).

Non-compliance with intervention

Successful implementation of the intervention depends on several factors: 1) Fidelity and dosage: are teachers implementing the intervention as intended according to the Research Lead? 2) Quality: are teachers aware of, and providing, feedback that adheres to the effectiveness quality according to the Research Lead? 3) Targeting: are teachers using the diagnostic tools to target delayed feedback at pupils with identified needs according to the Research Lead? We will take Factor 1 (Fidelity and dosage) measures to operationalise compliance with the intervention, whereas factors 2 (quality) and 3 (targeting) will be analysed as part of the process evaluation.

[8] This would be assessed at looking at the distribution of outcomes pre/post for treatment and control groups and comparing statistical parameters of dispersion.

We will explore the effect of the intervention on classes that were allocated to the intervention group and also implemented the intervention (compliers). The best measure of compliance will be data from the digital feedback app, more specifically data that indicate how often the app has been accessed by the teacher (i.e. a date-timestamp). *For each class, we will use the average number of times the app has been accessed per day (counting term-dates only) as a rate per 25 pupils (so that schools with smaller or larger classes are standardised).*

Other compliance measures we will collect include:

- Data from weekly coaching logs/records on the use of digital feedback in the classroom (self-reported utilisation). Compliers will be defined as the use of diagnostic assessments *at least once a week* for at least 30 out of 39 weeks in the school year. ; and
- Self-reported usage of the app by teachers. One of the questions in the teacher survey is '*How often do you use the following two types of digital feedback? Once a day or more than once a day / several times per week / once a week / once every two or three weeks / once a month* ' for both immediate and delayed feedback; we will count classes as compliers if the teacher indicates he or she has used digital feedback at least once a week both in the midline and the endline survey.

We will derive a Complier Average Causal Effect (CACE) measure by computing the proportion of intervention classes that actually receive the intervention, as part of the total intervention group, and use this to weight the ITT effect. If app-data are available, intervention classes will be counted as compliers *if they have accessed the app at least once a week for at least 30 out of 39 weeks in the school year*. If we cannot get access to app-data, we will use the measures proposed above, i.e. if they are identified as compliers on the coaching records and on the self-report. We will then calculate proportions of the intervention group that achieved / did not achieve this threshold (the percent compliant [PC]).

This complier–rate, divided by the ITT effect, represents the CACE estimate, and this will be computed and presented alongside the main effects. The CACE is represented in the following formula:

$$CACE=ITTC=ITTALL/PC$$

With ITTC= intention-to-treat effects in the compliers; ITTALL= main intention-to-treat effects and PC=proportion of classes allocated to the treatment group who actually receive the intervention as defined above.

Report tables

We will report according to the EEF template.

References

Allison, P. D. (2012). Handling Missing Data by Maximum Likelihood. *SAS Global Forum 2012 Statistics and Data Analysis*, 1–21. https://doi.org/10.1042/CS20080236

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, *91*(434), 444–455. https://doi.org/10.2307/2291629

Dong, N. and Maynard, R. A. (2013). *PowerUp!*: A tool for calculating minimum detectable effect sizes and sample size requirements for experimental and quasi-experimental designs. *Journal of Research on Educational Effectiveness*, *6*(1), 24-67. doi:

10.1080/19345747.2012.673143

Glennerster, R., & Takavarasha, K. (2014). Running randomized evaluations : a practical guide. London: Princeton University Press.

Hutchison, D., & Styles, B. (2010). A guide to running randomised controlled trials for educational researchers, 1–65.

Siddle, J. (2015) Verbal and visual-digital feedback on creative writing improves progress rates compared to written feedback in rural primary schools – a preliminary study. Kyra Teaching School Alliance.

Skinner, E.A., Kindermann, T.A., Furrer, C.J. (2009) A Motivational Perspective on Engagement and Disaffection: Conceptualization and Assessment of Children's Behavioral and Emotional Participation in Academic Activities in the Classroom. Educational and Psychological Measurement. Vol 69, Issue 3

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis : an introduction to basic and advanced multilevel modeling*.