Statistical Analysis Plan

Evaluating the effectiveness of Eedi formative assessment programme (previously Diagnostic Questions) on raising attainment in mathematics at GCSE

Alpha Plus Consultancy and Manchester Metropolitan University

Principal investigator(s): Mr Andrew Boyle and Stephen Morris

Template last updated: March 2018

PROJECT TITLE	Evaluating the effectiveness of Eedi formative assessment programme (previously Diagnostic Questions) on raising attainment in mathematics at GCSE
DEVELOPER (INSTITUTION)	Eedi and Behavourial Insights Team
EVALUATOR (INSTITUTION)	Alpha Plus Consultancy and Manchester Metropolitan University
PRINCIPAL INVESTIGATOR(S)	Mr Andrew Boyle and Professor Stephen Morris
TRIAL (CHIEF) STATISTICIAN	Professor Stephen Morris
SAP AUTHOR(S)	Professor Stephen Morris, Mr Andrew Smith and Dr Zsolt Kiss
TRIAL REGISTRATION NUMBER	ISRCTN62362872, http://www.isrctn.com/ISRCTN62362872
EVALUATION PROTOCOL URL OR HYPERLINK	*https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/EEDI_Protocol_2018.05.02_FINAL.pdf*

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]	2018.11.29	Orginal statistical analysis plan

Table of contents

Introduction

This study seeks to examine whether exposure to Eedi formative question setting platform and marking system for mathematics, among pupils and teachers in secondary schools, raises attainment in mathematics at GCSE. It also seeks to understand whether teachers' workload for teachers using Eedi for mathematics instruction at Years 10 and 11 is reduced.

The intervention is an online question setting and diagnostic platform for instruction in mathematics at GCSE known as Eedi (previously known as Diagnostic Questions), developed by Eedi (https://www.eedi.co.uk/).

The extent to which exposure to Eedi raises pupil attainment in GCSE mathematics and reduces teacher workloads is to be assessed on the basis of results from a two-arm cluster randomized controlled trial with random allocation to intervention and control groups at the school level on a 1:1 basis. Pupils (and their teachers) in Year 10 starting in September 2018 in intervention schools will be exposed to Eedi for a period of 2 years (Years 10 and 11). Year 10 pupils at September 2018 in control schools will not be able to access Eedi during the trial. It is important to note that pupils in control schools will have access to other similar online formative assessment and feedback platforms. Such exposure in the control group schools is considered to be business as usual activity. Therefore, the purpose of the analysis described in this SAP is to provide estimates of the average effects of intention to treat, in terms of controlled exposure to Eedi, on attainment at GCSE mathematics and teacher workloads, over business as usual.

Design overview

The following table provides a summary of the trial design. The trial is a two-arm cluster randomized controlled trial, where schools are assigned at random to intervention and control conditions on a 1:1 basis. The randomisation is stratified by region and batch. Stratification by batch occurs because schools were randomised in separate groups at three points in time The initial trial protocol stated that randomization would be conducted within two batched of schools entering the trial[1]. Due to anticipated delays in recruitment and to avoid further delaying training for participating schools, randomisation was conducted in three batches. The primary outcome is attainment in mathematics at GCSE among pupils within range of the intervention, as measured in the National Pupil Data base. The secondary outcome is a measure of self-reported teacher workload obtained from an online survey administered to intervention and control group teachers via email.

Trial type and number of arms		Two arm cluster randomised controlled trial
Unit of randomisation		School
Stratification variables (if applicable)		Region and batch
Primary outcome	variable	GCSE Mathematics
	measure (instrument, scale)	KS4_EBPTSMAT_PTQ_EE – Points score in maths Ebacc pillar
Secondary outcome(s)	variable(s)	Teacher workload hours per week
	measure(s) (instrument, scale)	Total time spent in a reference week preparing, setting, marking, recording and giving feedback related to mathematics home work (self reported per teacher). Recorded in hours and minutes

[1]https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/EEDI_Protocol_2018.05.02_FINAL.pdf

Sample size calculations overview

At the time of writing recruitment and randomisation of schools was complete. Minimum detectable effect sizes for the trial at protocol and randomisation are reported in the following table, for the full sample and those sample members 'ever in receipt' of free school meals (EverFSM6_p).

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
MDES		0.168	0.170	0.180	0.183
Proportion of Variance explained by Pre-Test	level 1 (pupil)	0.50	0.50	0.50	0.50
	level 2 (class)	0.00	0.00	0.00	0.00
	level 3 (school)	0.25	0.25	0.25	0.25
Intracluster correlations (ICCs)	level 2 (class)	0.05	0.05	0.05	0.05
	level 3 (school)	0.20	0.20	0.20	0.20
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 - class		24	8	30	7
Average cluster size - school		168	56	183	41
Number of schools	intervention	90	90	79	79
	control	90	90	79	79
	total	180	180	158	158
Number of pupils	intervention	15,120	5,040	14,856	3,384
	control	15,120	5,040	14,074	3,092
	total	30,240	10,080	28,930	6476

The sample of schools was obtained from records held by three major exam boards in England, Eedi itself and the Education Endowment Foundation. From these records schools were approached in order to be recruited to the trial. To be approached, schools had to be a non-selective secondary school and deemed *not* to be an existing active or extensive user of the Eedi platform. Schools considered for inclusion were those schools found to have 30 or fewer students that had accessed the Eedi platform and comleted quizzes between September 2017 and February 2018, taking into account total school size. Of 734 school records obtained, 458 met these trial inclusion criteria. All 458 schools were invited to 'book' a telephone call with the developers during which the trial was explained to them and during which they were invited to join the study.

A total of 287 telpehone calls were held between the developers and schools from early spring to mid-summer 2018 . Resulting from these calls, 190 schools signalled their intention to take part in the study through signing a memorandum of understanding with the developers and evaluators. Subsequent to signing the memorandum, 21 schools withdrew from the study for a variety of reasons and 11 failed to provide the required baseline data. The baseline data required were: the school

URN, the unique pupil number for all pupils in Year 9 in the school year 2017/18 (entering Year 10 from September 2018), their sex, date of birth and whether the child had ever qualified for free school meals, and whose parents had not withdrawn the child from the study. As a result, 158 schools were randomised in three batches as schools signed MoUs and supplied the necessary data. These batches were of size: 93 (randomised on 7th June 2018), 54 (29th June 2018) and 11 (20th July 2018) respectively.

Sample size calculations presented in the table above are those provided in the protocol which was published on 3rd May 2018 prior to randomisation and those based on the sample at randomisation (20th July 2018). In consultation with developers and EEF, the decision was taken at the time of writing the first draft of the trial protocol, to power the trial to detect an effect size that is a relatively modest one (see Hattie, 2008 for a discussion on effect sizes in education), due to the low cost associated with the intervention, and therefore the low effect size that is likely to be of importance. Prior expectations concerning the likely effect size that is minimally relevant had to be considered alongside the costs of powering the trial to identify a relatively small effect size (these costs are mainly those associated with recruitment of schools, training of teachers and collection of primary data). As a result of such considerations, at protocol, an effect size of 0.168 (standardised mean difference) was judged to be reasonable and consistent with an optimal trade off between the costs of recruitment and a modest minimally important effect size.

For the sample size calculations above, the values used for the intra-class correlation coefficients at levels 2 and 3 and proportions of variance explained at various levels through the inclusion of a covariate (KS2 fine grade score mathematics) were obtained from previous similar EEF-funded trials and studies that report summaries of intra-class correlation coefficients for education trials (Allen, Jerrim, Parameshwaran, & Thomson, 2018; Bloom, 2006; Bloom, Richburg-Hayes, & Black, 2007; Education Endowment Foundation, 2013; Hedges & Hedberg, 2013). For example, previous research commissioned by EEF has looked at correlations between GCSE and KS2 fine grade scores (Education Endowment Foundation, 2013). Similarly Allen et al. (2018) report intra-class correlation coefficients for KS4/GCSE outcomes obtained from NPD.

Due to difficulties in recruiting schools the final achieved sample size at randomization was 158 schools. This resulted in a modest increase in the minimum detectable effect size from 0.168 to 0.180. Althought less than ideal, this decrease in design sensitivity was not judged to undermine the trial to any significant extent.

Analysis of the 2015/16 school census suggests that roughly a third of pupils in state secondary schools had received free school meals in the past 6 years. This equates to roughly eight pupils per class. Maintaining the same assumption for the calculation of effect sizes under the primary analysis by FSM status, reveals that a trial involving the random allocation of some 180 schools to intervention and control conditions is consistent with a minimum detectable effect size of 0.170 for ever-free school meals subgroup analysis. This calculation was undertaken at the point in time the first draft of the study protocol was published. At randomization, and as discussed above, our sample of schools was 28 schools fewer than had been the target. Furthermore, the average number of pupils per class that were found to be ‘ever-free school meals’ was fewer than anticipated – six pupils as opposed to eight. These smaller sample sizes, in terms of the total number of schools and the lower number of pupils ever-free-school-meals per class, means that the minimum detectable effect size at randomization was 0.183 rather than 0.170 at protocol.

Analysis

Primary outcome analysis

The primary analysis will provide an estimate of the average effect of exposure to the intervention on attainment at GCSE on an intention to treat basis. The primary outcome is therefore attainment in mathematics at GCSE measured at the pupil level. Attainment in mathematics at GCSE is obtained for the study sample by linking study pupil trial records to the National Pupil Database.

Adopting attainment at GCSE as the primary outcomes has a number of advantages. First, considerable resources are devoted by exam boards to the writing and validation of GCSE questions. Second, the costs of collecting pupil level GCSE results are low compared to commercial standardised tests of attainment, given that results are extracted directly from National Pupil Database. Third, unlike administering separate standardised assessments of mathematics, using GCSE score at the primary outcome imposes no additional data collection burden on schools. Fourth, as a measure it is also less affected by loss to follow-up. Fifth, GCSE is widely recognised by employers, the government, colleges and universities and determines progression in education and students' future opportunities. GCSE grades are well understood so that results showing that an intervention has an effect in terms of GCSE grade is clear to, and interpretable by, stakeholders.

On the other hand, as Baird, et al. (2013) point out in their review of the evidence in connection to the recent reform of GCSE; as a measure of attainment, GCSE suffers from the incentive created for teachers to 'teach to the test'. They also point to examples of research stretching back over many years highlighting the limitations of examinations in terms of their reliability and predictive validity (Black & Wiliam, 1998; Gipps, 1994; James & Chilvers, 2001; Wiliam, 2001); though they also note that such claims are contested in the literature. The GCSE curriculum and therefore examinations, particularly mathematics, are broad in their coverage and results are essentially still reported as grades that lack granularity. Despite these disadvantages the importance of success at GCSE as a means of advancement and the study's funder's commitment to tackling inequality in attainment at GCSE, combined with the relatively low costs of obtaining GCSE results led to its selection as the primary outcome for this trial.

GCSE results are recorded in the NPD at the pupil level. This implies a hierarchical data set: pupils are nested within classes and classes within schools. To reflect this, a multi-level model is specified for the primary analysis with random effects at the school and class levels. The primary analysis is an adjusted analysis with covariates included in the model for prior attainment at the pupil level and variables used for stratification in the randomisation. The model is as follows:

$$Y_{ijk} = \beta_0 + \beta_1 T_i + \beta_2 X_{ijk} + \beta_3 S_i + \theta_i + \delta_{ij} + \varepsilon_{ijk} \ldots.[1]$$

Where 'i' indexes for school, 'j' class and 'k' pupil. Y_{ijk} is the GCSE mathematics score for pupil 'k' in class 'j', located within school 'i' as observed in the National Pupil Database subsequent to national examinations held in the summer of 2020. The variable T_i is a binary indicator coded one if the sample school is an intervention school and zero otherwise; X_{ijk} is a covariate that captures prior attainment in the form of the sample pupil's point score in their Key Stage 2 mathematics assessment obtained also from the National Pupil Database. The variable S_i is a composite stratification variable that captures the region in which the sample school is located and the batch in which the school was randomised. A sample estimate of the parameter β_1 is the estimated average effect of intention to treat in terms of exposure to Eedi and β_1 is therefore the model parameter of interest.

The parameters in Model 1 will be estimated using the multilevel mixed effects linear regression command 'mixed' in STATA v15.1 on the basis of maximum likelihood. In implementing this model we make the following assumptions concerning how the data are distributed in the population: 1) that pupil level test scores Y_{ijk} are normally distributed around class means with constant variance σ^2_{WC} ; 2) that class room level mean test scores are normally distributed around population school means with constant variance σ^2_{BC}; and 3) that school means are normally distributed around intervention and control group means with constant variance σ^2_{BS} (Hedges, 2011).

The within treatment groups total variance is therefore composed of the sum of the within class, between class and between school variances as follows:

$$\sigma^2_{WT} = \sigma^2_{BS} + \sigma^2_{BC} + \sigma^2_{WC}$$

Two intra class correlation coefficients can be defined at the school and class levels:

$$\rho_S = \frac{\sigma^2_{BS}}{\sigma^2_{BS}+\sigma^2_{BC}+\sigma^2_{WC}} = \frac{\sigma^2_{BS}}{\sigma^2_{WT}} \ldots.[1A]$$

$$\rho_C = \frac{\sigma^2_{BC}}{\sigma^2_{BS}+\sigma^2_{BC}+\sigma^2_{WC}} = \frac{\sigma^2_{BC}}{\sigma^2_{WT}} \ldots\ldots[1B]$$

Estimates of the intra class correlation coefficients will be presented for the primary analysis based on results obtained from fitting Model [1].

We propose to estimate three further specifications as sensitivity checks in the primary analysis (in addition to these specifications further analyses is proposed below to explore the sensitivity of estimates to missing data):

1) An unadjusted model (Model [2]) to sensitivity check the stability of estimated effects to the removal of the covariate capturing prior attainment. The model is of the following form:

$$Y_{ijk} = \beta_0 + \beta_1 T_i + \theta_i + \delta_{ij} + \varepsilon_{ijk} \ldots [2]$$

2) A model (Model [3]) including a term capturing the mean KS2 points score in mathematics for the school along side a term capturing a school centred pupil level points score at KS2 of the following form:

$$Y_{ijk} = \beta_0 + \beta_1 T_i + \beta_2 (X_{ijk} - \bar{X}_i) + \beta_3 S_i + \beta_4 \bar{X}_i + \theta_i + \delta_{ij} + \varepsilon_{ijk} \ldots..[3]$$

This specification enables us to determine whether the inclusion of a measure of school level average prior attainment for the sample improves the precision of the estimates and how far estimates are sensitive to the inclusion of such a covariate.

3) Third, a fuller specification (Model 4) including a range of further covariates obtained from pupil records supplied to the intervention developers by schools will be estimated. These covariates will be sex, month of birth and ever-free school meals:

$$Y_{ijk} = \beta_0 + \beta_1 T_i + \beta_2 X_{ijk} + \beta_3 S_i + \beta_4 SEX_{ijk} + \beta_5 MoB_{ijk} + \beta_6 FSM_{ijk} + \theta_i + \delta_{ij} + \varepsilon_{ijk} \ldots..[4]$$

The assumptions set out above for Model 1 in terms of normality and constant variance are assumed to hold for the data in relation to the estimation of Models 2, 3 and 4. For Models 2, 3 and 4 we will report sample estimates for the intra class correlation coefficients described in models 1A and 1B above for post-intervention outcomes. Parameter estimates for Models 2, 3 and 4 will be obtained through implementing the multilevel mixed effects linear regression command 'mixed' in STATA v15.1 on the basis of maximum likelihood.

Secondary outcome analysis

The secondary outcome for this trial is teachers' mathematics related workload. The purpose of the secondary analysis is to examine whether teachers' exposure to Eedi reduces their workload relative to business as usual. The analysis will be conducted on an intention to treat basis comparing average workload among intervention group teachers with that recorded by control group teachers.

In order to place the analysis in context, we discuss briefly here how teacher workload data are generated. An online questionnaire was administered to teachers in control and intervention schools prior to commencement of the intervention (summer term 2018), but in a number of cases subsequent to randomisation, asking respondents teaching mathematics to both Year 10 and 11 pupils to report the number of hours and minutes in a reference week they spent on the following tasks:

- Preparing maths homework
- Setting maths homework
- Marking maths homework
- Recording, chasing and analysing maths homework
- Giving verbal feedback on maths homework to pupils

- Planning maths lessons
- Communicating with parents and carers regarding maths performance

The sample of teachers selected for the survey was obtained by the intervention developers from schools participating in the study. The sample consisted of teachers' email addresses. In total the sample contained teachers from 143 of the 158 schools randomised. Across these 143 schools a sample of 1663 teacher email addresses was identified. Questionnaires were distributed to teachers at the end of June/beginning of July 2018. Of the 1663 email addresses 87 contained errors and were undeliverable. 137 teachers indicated that the survey was not relevant to them. In total 686 questionnaires were returned, 352 from intervention group teachers and 327 from control group teachers, representing a total response rate of 48 per cent (in 143 schools).

The survey questionnaire will be re-administered to Year 10 teachers in December 2018, March 2019 and Year 11 teachers in March 2020. The secondary outcome - 'teacher workload' – will be constructed from data at all four waves of survey data collection in a consistent manner by adding up the amount of time spent on each of the tasks set out above to provide a measure of total workload in the reference week measured in minutes. Total workload will form the dependent variable in the statistical analysis.

Analysis of the secondary outcome will commence with a descriptive assessment of simple mean workload by intervention and control groups, disaggregated by component tasks outlined above. We propose that three models are estimated on these teacher workload survey data where total workload is the dependent variable.

1) The first of these models will be estimated on three subsets of the data:
   - Those Year 10 teachers responding to the survey at December 2018 (post-intervention) and at July 2018 (pre-intervention)
   - Those Year 10 teachers responding to the survey at March 2019 (post-intervention) and at July 2018 (pre-intervention); and
   - Year 11 teachers responding to the survey at March 2020 (post-intervention) and Year 11 teachers responding at July 2018 (pre-intervention)

   Thus there will be three separate adjusted estimates of the effect of exposure to Eedi on teachers workloads at December 2018, March 2019 and March 2020 obtained from a model specified as follows:

$$Y_{ij} = \beta_0 + \beta_1 T_i + \beta_2 X_{ij} + \beta_3 S_i + \beta_4 NTQ_{ij} + \beta_5 Spec_{ij} + \theta_i + \varepsilon_{ij} \ldots\ldots [5]$$

   Each subset of the data will be a balanced panel, in that each teacher in the sample will provide an estimate of workload at two measurement occasions (pre and post-intervention). There are therefore some obvious limitations to this approach steming from sample attrition between the two measurement points which we address through proposing further specifications below to assess the sensivity of results to these restrictions placed on the sample.

   As workload is measured at the teacher level and teachers are nested within schools a two level hierarchical linear model with school-level random effects is proposed (Model [5] above). Here 'i' indexes for school and 'j' teacher. Y_{ij} is a measure of post-intervention teacher workload for teacher 'j' in school 'i'. The variable T_i is a binary indicator coded one if the sample school is an intervention school and zero if a control school and S_i the variables region and batch used in stratification. X_{ij} is a pre-intervention measure of workload for teacher 'i' collected, as discussed, prior to the commencement of the intervention in the summer term 2018. The variables NTQ_{ij} and $Spec_{ij}$ capture whether teacher 'j' in school 'i' was a newly qualified teacher and/or a mathematics specialist teacher at the summer term 2018. The parameter β_1 represents the average effect of intention to treat of exposure to Eedi on average teacher workloads and is the parameter of interest. The parameters in the model

will be estimated using the multilevel mixed effects linear regression command 'mixed' in STATA v15.1 on the basis of maximum likelihood. The assumptions of normality and constant variance apply.

2) The second specification we propose for estimating the average effects of Eedi on teacher workload simply compares separately average responses at the three measurement occasions post-intervention in intervention and control groups. Thus again we propose to estimate three models on survey data collected at December 2018 (Year 10), March 2019 (Year 10) and March 2020 (Year 11) but we do not include a covariate capturing pre-intervention workload at July 2018 in the analysis. In this sense the analysis can be considered an unadjusted analysis. The model takes the following form:

$$Y_{ij} = \beta_0 + \beta_1 T_i + \beta_2 S_i + \theta_i + \varepsilon_{ij} \ldots [6]$$

Y_{ij} represents the measure of post-intervention work load at the appropriate measurement occasion for teacher teacher 'j' in school 'i'. The variable T_i is a binary indicator coded one if the sample school is an intervention school and zero if a control school and S_i the variables region and batch used in stratification.

3) The final specification is more flexible than those discussed thus far. It permits us to estimate the average treatment effect using all the information collected from teachers at each measurement occasion. Here teachers can supply a number of observations on workload where measurement occasions are not equally spaced in time. Measurement occasions are nested within teachers and teachers within schools and therefore we propose a repeated measures multi-level model (Hox, Moerbeek, & Van de Schoot, 2018).

The simpliest form of this model contains a variable indicating the measurement occasion at level 1 P_{ijt} and the binary indicator T_i at the school level – level 3 (in practice the model will also contain an indicator for the year group the teacher supplying the observation is teaching which we omit for simplicity). The model can be written in the following form by levels:

$$Y_{ijt} = \pi_0 + \pi_1 P_{ijt} + \varepsilon_{ijt}$$

$$\pi_0 = \pi_{00} + \delta_{0ij}$$

$$\pi_1 = \pi_{10} + \delta_{1ij}$$

$$\pi_{00} = \beta_{000} + \beta_{001} T_j + \theta_{0i}$$

$$\pi_{10} = \beta_{010} + \beta_{011} T_j + \theta_{1i}$$

Through substitution we end up with the following model from which we will obtain the sample estimate of the average effect of exposure to Eedi on teacher workloads over time:

$$Y_{ijt} = \beta_{000} + \beta_{001} T_j + \beta_{010} P_{ijt} + \beta_{011} T_j P_{ijt} + (\theta_{1i} P_{ijt} + \delta_{1ij} P_{ijt} + \theta_{0i} + \delta_{0ij} + \varepsilon_{ijt})$$

In this model separate linear time trends are estimated for intervention and control groups relative to the respective starting positions pre-intervention. The parameter of interest is β_{011}. The main difference between this model and those specified previously is that we now allow parameters for the fixed terms in the model to vary (the slopes now vary as well as the intercepts). The complex error term is in parantheses. As with the models described in the previous sections, sample estimates will be obtained through implementing the multilevel mixed effects linear regression command 'mixed' in STATA v15.1 on the basis of maximum likelihood.

Interim analyses

No interim analysis is proposed.

Subgroup analyses

Two subgroup analyses are specified on the basis of the primary outcome. These are 1) a specification that examines the effects of Eedi by whether sample pupils were ever in receipt of free school meals (as recorded in the data supplied by the school and made available to us by the developers); and 2) by sex, that is whether effects vary for boys and girls (where sex is recorded in the data supplied by school and made available to us by the developers).

As with the primary analysis attainment is measured at the pupil level and pupils are nested within classes and classes within schools. Therefore a multilevel model is specified of the following form with respect for the analysis based on free school meals:

$$Y_{ijk} = \beta_0 + \beta_1 T_i + \beta_2 X_{ijk} + \beta_3 S_i + \beta_4 FSM_{ijk} + \beta_5 (T_i * FSM_{ijk}) + \theta_i + \delta_{ij} + \varepsilon_{ijk}$$

Sample estimates of β_5 repesent the effect for those pupils that have been in receipt of FSM in the intervention group. A similar specification will be estimated based on pupil sex as follows, with analogous interpretation:

$$Y_{ijk} = \beta_0 + \beta_1 T_i + \beta_2 X_{ijk} + \beta_3 S_i + \beta_4 SEX_{ijk} + \beta_5 (T_i * SEX_{ijk}) + \theta_i + \delta_{ij} + \varepsilon_{ijk}$$

Here the variable SEX is coded '1' if the pupil is male and zero otherwise. Both specifications will be estimated using the multilevel mixed effects linear regression command in STATA v15.1 on the basis of maximum likelihood with standard assumptions for normality and constant variance.

Additional analyses

A non-experimental analysis is proposed to examine the association between the number of quizzes a pupil attempts and their subsequent result at GCSE mathematics. This analysis will be performed on the intervention group sample only, for whom the developers are able to collect participation data.

$$Y_{ijk} = \beta_0 + \beta_1 Q_{ijk} + \beta_2 X_{ijk} + \beta_3 S_i + \theta_i + \delta_{ij} + \varepsilon_{ijk}$$

Again pupils are nested within classes and classes with schools. Here Q_{ijk} represents the number of quizzes attempted by student 'k' in class 'j' and school 'i' over the life of the trial. The sample estimate of β_1 will represent the association between the number of quizzes attempted and GCSE attainment. All other apects of this analysis will remain as described previously.

Imbalance at baseline

Despite randomising treatments, imbalances between intervention and control groups occur. These imbalances arise from the randomisation procedure (the estimated treatment effect differs from the population treatment effect as a result of chance variations) and sample attrition processes that may differ in intervention and control samples. Imbalances arising from randomisation are captured in the standard error and relate to precision (standard errors can be attenuated through the inclusion of baseline covariates such as pre-test attainment scores making estimates more precise), whereas sample attrition can lead to both losses in precision and bias.

In order to deterime the extent to which imbalances in intervention and control samples are present, we will conduct balance tests on both the 'as randomised' and 'as analysed' samples as an adjunct to the primary analysis. These balance test will compare the characteristics of intervention and control groups in terms of their baseline characteristics on the basis of the observable measures we have collected in both the 'as randomised' and 'as analysed' samples. The balance tests will take the forms of tabular analyses comparing means and proportions of observable characteristics in intervention and control samples, at both individual pupil and school levels in the data.

For the sample 'as randomisation', the following variables will be included in the tabular analysis, these are variables available for all cases regardless of whether records are linked successfully to the National Pupil Database, and were provided to the evaluators by the developers

- Sex (pupil level)
- Ever-FSM (pupil level)
- Month of birth (pupil level)
- Proportion of records linked successfuly to NPD (pupil level)
- School type (School level)
- School size Year 9 (School level)
- Region (School level)
- LAD (School level)
- Exam Board (School level)

Tabular analysis will present means and proportions in intervention and control samples, raw and standardised mean differences.

For the 'as analysed' sample comprising the linked trial/National Pupil Database data set, comparisons between intervention and control groups can be made on the basis of the following variables in addition to those mentioned above:

- Points score at KS2 mathematics
- Average points score at KS2 mathematics at the school level

Again tabular analysis will present means and proportions in intervention and control samples, raw and standardised mean differences.

Missing data

The primary analysis is dependent on data collected at randomisation by the developers direct from schools (these are potential covariates in the proposed treatment effects models) and KS2 and GCSE scores obtained from linking trial records to the National Pupil Database (these are prior attainment and the post intervention outcome measure). At analysis, it is most likely that missing data will stem from (1) schools withdrawing from the study subsequent to randomisation and asking that their pupil records are not linked to the National Pupil Database from which the primary outcome and measure of prior attainment is obtained; and (2) failure to link trial records to the NPD due to errors in the recording of UPNs. Given the data available to us we do not anticipate missing data on covariates to be a significant issue. Furthermore, it seems reasonable to suppose that recording errors are *unlikely* to differ systematically between intervention and control groups.

In order to assess the extent to which any missing data on the primary outcome is likely to be problematic, we will first examine whether more than five per cent of pupil records are missing in the 'as analysed' compared to the 'as randomised' samples. If we find attrition rates in excess of this threshold and/or substantially different rates of loss in intervention and control groups we propose to model the probability of the outcome variable being missing at the pupil level in the form of a logistic regression model containing the following covariates (the so called drop out model):

At the pupil level:

- Sex (pupil level)
- Ever-FSM (pupil level)
- Month of birth (pupil level)

And at the school level:

- % of Year 9 cohort (Summer 2018) ever-free school meals
- School size – Year cohort (Summer 2018)
- School type
- Region
- LAD
- Exam Board

This model will correct for the clustering of cases within classes and schools. The procedure 'melogit' in STATA v15 will be used for this purpose. Where there is evidence that individual covariates in the model appear to be associated with missing values on the dependent variable then these variables will be added as covariates to those in model [4] and the adjusted analysis re-run in order to assess the susceptibility of the primary estimates to potential biases resulting from missingness. This approach will enable us to recover unbiased estimates of the average effect of intention to treat in the presence of missing data on the outcome where missing data processes are 'missing at random'. However, the analysis will be performed on the completed cases data set. Thus on a reduced sample size and thereby raising the Type II error rate.

Moreover, our assumption is also that due to the processes by which data are obtained via schools at randomisation, we will not encounter missing values on covariates. If the scale of missing observations on the outcome and/or the rate of missing data on covariates appears to be substantial (where the percentage of cases with missing values on a given covariate is greater than five to ten per cent for example), such that in the case of the former statistical power is substantially reduced, or in the case of the latter potential bias arises and reduced statistical power, the use of multiple imputation will be considered using the 'mi' suite of commands in STATA v15. We propose to conduct multiple imputation (up to five imputations) on the basis of 100 interations and making the standard multivariate normal assumptions. The logistic regression model described above will be used to determine the auxiliary variables used in the multiple imputation.

Compliance

Schools/teachers and to some extent pupils may fail to comply with the intervention conditions to which they were assigned. Schools assigned to the intervention may subsequent to randomisation choose not to use the Eedi platform. Teachers in intervention schools are also likely to be able to act with some autonomy and choose whether to use Eedi or not. Choices made by schools and teachers are likely to have a decisive effect on compliance at the pupil level. Among control schools, those schools that choose to leave the trial will be able to access Eedi. Thus teachers will be able to create accounts and pupils complete quizzes. As a result, patterms of compliance/non-compliance in this study involve complex interactions between schools, classroom teachers and their pupils.

In many trials intention to treat estimates on the primary outcome are adjusted for non-compliance (Gerber, Alan & Green, Donald, 2012). This adjustment is made to estimate a different parameter to intention to treat, usually referred to as the complier average causal effect (CACE). In the case of this present study there is the potential for non-compliance in both intervention and control samples. In order to simplify the problem of how to adjusted intention to treat estimates for non-compliance we make the observation that an estimate of compliance among the as randomised sample at the pupil level would in effect capture compliance at all levels in the data. Thus we define an intervention pupil as being non-compliant if over the course of the study they fail to attempt a single quiz. Likewise, we define a control group pupil as being non-compliant if they sit or attempt at least one quiz. Administrative systems desiged to manage the Eedi platform enable us to construct measures of compliance on this basis and link such measures to the trial sample.

With these data at our disposal we propose to estimate the average treatment effect on compliers (the complier average causal effect or CACE) as follows for the primary analysis:

$$CACE_Y = \frac{\widehat{\beta_1}}{P_I - P_C} \ldots.[7]$$

Where $\widehat{\beta_1}$ is the sample estimate of β_1 in model [1] (that is the sample estimate of intention to treat based on completed cases assuming ignorable missingness), and P_I is the proportion of pupils attempting at least one quiz in the intervention group and likewise P_C the proportion of control group pupils attempting at least one quiz. The denominator in [7] can also be used to adjust the standard error obtained from fitting model [1] to the data. The assumptions made in peforming this calculation, such that it provides an unbiased estimate of the average treatment effect for compliers are: a) we rule out the possibility of defiers (people who use Eedi only when assigned to control conditions – the

monotonicity assumption); and b) that the intervention has no impact on those who do not use it – that is on those that complete no quizzes.

Effect size calculation

The Education Endowment Foundation require that results from the primary analysis are reported as effect sizes and specifically as 'Hedges g' (Hedges, 1981). We propose, based on models [1] and [2], to derive a sample estimate of the effect size equivalent to Hedges g with 95 per cent confidence interval consistent with the requirements of EEF in the following way. Hedges g, with three levels of clustering the data and unequal sample sizes is defined as:

$$\hat{\Delta}_g = \frac{\widehat{\beta_1}}{S_{WT}} \sqrt{1 - \frac{2(p_u - 1)\rho_s + 2(n_u - 1)\rho_c}{N-2}} \ldots\ldots [8],$$

Where $\widehat{\beta_1}$ is the adjusted mean difference in attainment in GCSE mathematics between intervention and control groups obtained from fitting model [1] to the sample data. S_{WT} is the within group pooled standard deviation. It is based on the unconditional rather than conditional sample variance and therefore calculated direct from the sample data as follows (Hedges, 2011)

$$S_{WT}^2 = \frac{\sum_{i=1}^{m^I} \sum_{j=1}^{p_i^I} \sum_{k=1}^{n_{ij}^I} (Y_{ijk}^I - \bar{Y}_{\circ\circ\circ}^I)^2 + \sum_{i=1}^{m^C} \sum_{j=1}^{p_i^C} \sum_{k=1}^{n_{ij}^C} (Y_{ijk}^C - \bar{Y}_{\circ\circ\circ}^C)^2}{N-2}$$

Where 'm^I' is the total number of schools in the intervention sample, 'p_i^I' the total number of classes and 'n_{ij}^I' the total number of pupils, likewise in the control group. $\bar{Y}_{\circ\circ\circ}^I$ is the mean outcome among intervention schools and $\bar{Y}_{\circ\circ\circ}^C$ the mean outcome among control schools. The final term in equation [8] adjusts for the fact that the sample data are clustered (see equations 1A and 1B) and noting that:

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

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

As Hedges (2011) explains inference is complex in the situation where class sizes and the numbers of classes per school vary (as is the case in our sample), and the most tractable parameteric solutions proposed are approximations, though conservative.

Our approach is to construct confidence intervals for the sample estimate of Hedges g based on the bootstrap. This will involve sampling from the residuals resulting from fitting model [2][2] using the 'runmlwin' command in STATA. This approach effectively enables us to use the bootstap procedures available in the software package MLWin 3.02 within the STATA v15 operating environment. From the resulting empirical distribution function we wil calculate Hedges g on each run, of which we intend 1,000. On the basis of these results confidence intervals can be obtained through following the general principles and steps set out in Mooney, Duval, & Duvall (1993).

References

Allen, R., Jerrim, J., Parameshwaran, M., & Thomson, D. (2018). *Properties of commercial tests in the EEF database*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Publications/EEF_Research_Papers/Research_Paper_1_-_Properties_of_commercial_tests.pdf

Baird, J.-A., Ahmed, A., Hopfenbeck, T., Brown, C., & Elliott, V. (2013). *Research evidence relating to proposals for reform of the GCSE*. Oxford: Oxford University Centre for Educational

[2] model [2] is used in the re-sampling procedures because it will produces an unconditional empirical sampling distribution for 'Hedges g'.

Assessment. Retrieved from http://content.yudu.com/Library/A24v28/Researchevidencerela/resources/index.htm?referrerUrl=http%3A%2F%2Ffree.yudu.com%2Fitem%2Fdetails%2F837575%2FResearch-evidence-relating-to-proposals-for-reform-of-the-GCSE

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74.

Bloom, H. S. (2006). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytical approaches* (pp. 115–171). New York, NY: Russell Sage Foundation.

Bloom, H. S., Richburg-Hayes, L., & 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. http://doi.org/10.3102/0162373707299550

Education Endowment Foundation. (2013). *Pre-testing in EEF evaluations*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Protocol_or_SAP/Pre-testing_paper.pdf

Gerber, Alan, S., & Green, Donald, P. (2012). *Field experiments: Design, analysis, and interpretation*. New York, NY: W. W. Norton & Company.

Gipps, C. (1994). Developments in Educational Assessment: what makes a good test? *Assessment in Education: Principles, Policy & Practice*, *1*(3), 283–292.

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. routledge.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128.

Hedges, L. V. (2011). Effect sizes in three-level cluster-randomized experiments. *Journal of Educational and Behavioral Statistics*, *36*(3), 346–380.

Hedges, L. V, & Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning two-and three-level cluster-randomized experiments in education. *Evaluation Review*, *37*(6), 445–489.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2018). *Multilevel Analysis: Techniques and Applications* (3rd ed.). NewYork: Routledge.

James, D., & Chilvers, C. (2001). Academic and non-academic predictors of success on the Nottingham undergraduate medical course 1970–1995. *Medical Education*, *35*(11), 1056–1064.

Mooney, C. Z., Duval, R. D., & Duvall, R. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. Sage.

Wiliam, D. (2001). Reliability, validity, and all that jazz. *Education 3-13*, *29*(3), 17–21.