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

Evaluation Study Plan

Evaluator (institution): AlphaPlus Consultancy and Manchester Metropolitan University

Principal investigator(s): Andrew Boyle and Stephen Morris

PROJECT TITLE	Evaluating the effectiveness of Eedi formative assessment programme (previously Diagnostic Questions) on raising attainment in mathematics at GCSE
EVALUATOR (INSTITUTION)	AlphaPlus Consultancy and Manchester Metropolitan University
PRINCIPAL INVESTIGATOR(S)	Mr Andrew Boyle and Professor Stephen Morris
STUDY PLAN AUTHOR(S)	Professor Stephen Morris and Dr Zsolt Kiss
STUDY DESIGN	Cluster randomised controlled trial with 158 school initially assigned at random to intervention and control groups. GCSE mathematics attainment for 2020 cohort was to be the primary outcome. Following Covid-19 outbreak only the original trial secondary outcome - a measure of teacher workload – is available for analysis due to the cancellation of national examinations in 2020. Moreover, high levels of non-adherence and differential loss-to-follow-up across trial arms leads to teacher workload data being analysed non-experimentally using inverse probability weight regression adjustment. As a result, all analyses discussed in this protocol is considered exploratory.
PUPIL AGE RANGE AND KEY STAGE	KS4 but the analysis is conducted on a sample of teachers
NUMBER OF SCHOOLS	149
NUMBER OF TEACHERS	824
OUTCOME MEASURE AND SOURCE FOR EXPLORATORY IMPACT ANALYSIS	Mathematics related teacher workload obtained via teacher self-reported hours in survey questionnaires (hours per week)
OTHER OUTCOME MEASURES AND SOURCES	A range of additional exploratory analyses are proposed in which outcomes capture pupil take-up and use of the intervention among pupils in intervention schools

Study plan version history

VERSION	DATE	REASON FOR REVISION
3.0 [*latest*]	2020.11.15	Substantial revisions to the study plan required to reflect: • the absence of the primary outcome measure for this study due to the cancellation of GCSE summer examinations for 2020, • substantial non-adherence to the intervention; and • differential loss to follow-up found in teacher surveys from which outcome measures are derived. Further, due to the onset of the Covid-19 pandemic a range of exploratory observational analysis has been deemed desirable drawing on administrative data generated through the Eedi online platform.
2.0 [unpublished]	2020.02.28	Amendments to revise NPD variable as a measure of primary outcome and to clarify that analysis will be based on GCSE grade (as distinct from score). Changes to secondary outcome analysis to account for lack of availability of matched responses, and high non-response in successive survey rounds. Minor changes to specified sensitivity analysis for the primary outcome.
1.0 [*original*]	2018.11.09	Original statistical analysis plan

Table of contents

Background and study rationale

For a discussion of the theoretical and scientific background, policy context and rationale for this evaluation readers are referred to the original trial protocol that can be found on the Education Endowment Foundation's (EEF) website (Seymour & Morris, 2018)[1].

This study was originally designed as a pragmatic cluster randomised controlled trial. It involved the allocation of 158 secondary schools in England, at random, on a 1:1 basis, to intervention and control groups. Pupils entering Year 10 at September 2018 in intervention schools were to be exposed to the intervention for two years as were their teachers. Pupil attainment in mathematics GCSE was to form the study's primary outcome. Pupils in the focal cohort were due to sit their GCSEs in the summer of 2020. The Covid-19 pandemic led to English schools closing from the end of March 2020, with the exception of schooling for vulnerable children and the children of key workers. At the same time the cancellation of summer GCSE examinations was announced. The predicted grades that replaced actual examination results were not deemed suitable as a primary outcome. As a result, the primary outcome for this study was no longer available.

The initial evaluation design also specified a secondary outcome. This was to be weekly mathematics-related teacher workload, specifically in relation to homework, measured in self-reported hours for a given reference week. The measure of teacher workload was to be obtained from surveys of teachers teaching mathematics to the focal pupil cohorts prior to randomisation, at December 2018, March 2019 and March 2020.

A full statistical analysis plan that set out the proposed experimental analysis was published in November 2018 (Morris, Smith, & Kiss, 2018)[2]. This was updated with a revised statistical analysis plan in March 2020. The updated plan was not published due to the consequences of the Covid-19 pandemic rendering much the proposed analysis impossible. The evaluation design also comprised a full implementation process evaluation (IPE), which has been carried out and which to a large extent has been unaffected by Covid-19. Details of the design of the IPE can also be found in the original trial protocol (Seymour & Morris, 2018).

Surveys administered to teachers responsible for delivering mathematics teaching to children in the focal cohorts did proceed as planned. The secondary outcome, teacher homework-related workload in hours per week, was to be derived from responses to these surveys and analysed experimentally. Although the administration of surveys was largely unaffected by the pandemic a number of other problems suggest that the resulting data cannot be analysed straightforwardly as experimental data. Instead, we should treat the data as observational and analyse them using methods usually reserved for observational data / non-experimental data.

[1] https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/diagnostic-questions/

[2] Ibid.

Table 1: Achieved sample sizes by intervention and control schools – teacher survey

	Baseline	**Wave 1**	**Wave 2**	**Wave 3**
Intervention Group				
Total achieved sample of teachers	352	208	167	64
Teachers reporting use of Eedi	n/a	164	134	44
Total schools	66	59	48	24
Control Group				
Total achieved sample of teachers	327	304	281	181
Teachers reporting use of Eedi	n/a	0	0	0
Total schools	70	69	66	39
Totals				
Total achieved sample of teachers	679	512	448	245
Teachers reporting use of Eedi	n/a	164	134	44
Total schools	136	128	114	63

In summary, exploratory impact analysis is now proposed that looks at the the effects of the take-up of Eedi by teachers on their mathematics homework-related workload, where measures of self-reported workload for a reference week are derived from the teacher survey. A measure of teachers' use of the Eedi platform in a reference week is obtained from the teacher survey data . Further, a range of secondary and exploratory analysis is proposed. These are analysis of data from the Eedi administrative system which is used to manage the platform and interactions between teachers and pupils using the software in intervention schools. A range of analyses are proposed that make use of these data to explore how pupils used the system, and how take-up and use varied across groups of pupils that had access to the platform.

In relation to the analysis as now defined, there are a number of challenges faced, that necessitates an approach that does not rely on simple experimental contrasts:

- Among teachers in the intervention group there is evidence of substantial **non-adherence to the intervention**. Non-adherence varied between treatment and control schools as expected consistent with randomisation to the platform but it also varied between interventions schools and between teachers within the same intervention school;
- The achieved teacher survey sample is quite unbalanced by intervention and control group, which is likely to reflect **different loss to follow-up processes in the two arms** of the study. The proportion of the total sample in the intervention group at Waves 1-3, bearing in mind this should be roughly 50 per cent, stood at 41 per cent, 37 per cent and 26 per cent respectively; and
- The teacher survey questionnaires due to the **lack of a sampling frame** (which was due to be provided by the intervention developers) were distributed by schools to teachers in range of the study. This means that a teacher-level survey non-response weight cannot be derived for use in our analysis. This is clearly less than ideal and is an important limitation of the proposed analysis. By way of partial mitigation, we propose deriving a school-level non-response weight to correct for teacher workload observations missing from teachers in schools that did not participate in the survey.

Table 1 shows that at Waves 1 to 3 of the survey, 79 per cent, 80 per cent and 69 per cent of teachers that responded in the intervention group were using the Eedi system at the time of the survey. This means that the Eedi platform could have affected their reported workloads. Furthermore, we also find that adherence varies within intervention schools, and so is to some extent a feature of individual teacher behaviour rather than being wholly constrained by school level choice. For example, in Wave 3, on average there were 3.7 teachers who used the platform per school (there were 4.9 teachers per schools in total, on average). The standard deviation was 1.4, suggesting that there were differences in usage between schools. Moreover, in approximately 30% of intervention schools where we have a measure of uptake of the platform, between 25% to 80% of teachers used the platform.

As noted, given the initial design of the trial and in the absence of differential loss to follow-up, we would expect the teacher samples to be roughly equivalent in size in the intervention and control groups. Table 1 shows that this is not the case and that the balance between intervention and control groups in absolute sample size deteriorated over time. Moreover, these imbalances in sample size are likely to lead to imbalances in average values across covariates by intervention and control groups (both measured and unmeasured).

If a sample frame were available for teachers in range of this study, any covariates contained on the sampling frame could be used to generate survey weights that might adjust for measured differences in loss to follow-up in the available covariates and outcome measure. As mentioned previously no such sampling frame was made available to the evaluators.

The challenges noted here are significant and the strategies available less than ideal, even in non-experimental terms. Notwithstanding these concerns, we propose an analysis that seeks an estimate of the average treatment effect on the treated, of exposure to Eedi, on teacher workload from the achieved sample data. This analysis aims to adjust for non-response to the survey in the treatment and intervention arms of the study at the school-level subsequent to the baseline, and to adjust for imbalances in the achieved sample across teachers that do and do not use the Eedi platform. By taking these steps we hope to get an estimate closer to the true average effect of treatment on the treated, or at a minimum not make any biases in the data worse. However, fundamentally, we have only limited school-level information from outside of the achieved sample in order to improve our estimates.

This SAP also provides further discussion of analysis that will be conducted on the basis of data collected from the intervention group pupils, drawing primarily on administrative data derived from the Eedi platform. These data are entirely observational in nature. The research questions to be addressed in these analyses are specified below as are the regression models used in the relevant analysis.

Intervention

For a full description of the intervention, its logic model, rationale and the period of time over which the intervention operated within the context of this trial, readers are referred to the original trial protocol (Seymour & Morris, 2018). This can be found on the EEF's website, at https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/diagnostic-questions/

Research questions

This statistical analysis plan describes the approaches taken to addressing six questions; one of which relates to analyses of the teacher survey data and the other five to analysis of administrative data extracted from the Eedi management system and linked to data from the trial and National Pupil Database (NPD).

RESEARCH QUESTION

1. What is the difference in the average number of hours per week spent on mathematics related non-classroom work among teachers that report using the Eedi platform compared to teachers that do not report use the platform?
2. How does the number of quizzes ***set*** vary across the sample of children in schools allocated to the intervention by KS2 attainment in mathematics, gender, month of birth and free-school meals?
3. How does the number of quizzes ***started*** vary across the sample of children in schools allocated to the intervention by KS2 attainment in mathematics, gender, month of birth and free-school meals?
4. How does the number of quizzes ***completed*** vary across the sample of children in schools allocated to the intervention by KS2 attainment in mathematics gender, month of birth and free-school meals?
5. What is the average difference in the quiz scores for pairs of quizzes that test for the same subject knowledge at two points in time: right after teaching the lesson and three weeks later (while controlling for by KS2 attainment in mathematics, gender, month of birth and free-school meals)?
6. How does the usage of the Eedi platform vary during the Covid-19 lockdown across the sample of children in schools allocated to the intervention, by KS2 attainment in mathematics, gender, month of birth and free-school meals?

Design overview

Initially, this study was designed as a cluster randomised trial, involving the random allocation of 158 schools to intervention and control conditions, with attainment in national GCSE examinations in mathematics the primary outcome. As described above, it is no longer possible to use GCSE attainment as the primary outcome. A secondary outcome was to be teacher mathematics-related workload derived from teacher-self-reported-workload collected through a survey of teachers. The survey was last administered at Wave 3 just prior to the closure of schools due to the Covid-19 pandemic, and so can be analysed.

Although the samples of teachers are generated from an experiment, sample attrition and non-adherence leads to an achieved sample by use of Eedi that is likely to be imbalanced, and imbalanced to an extent beyond that due to randomisation alone. To adjust for these imbalances, as far as is possible, the proposed exploratory impact analysis consists of treatment effect estimates obtained through inverse probability weighted regression adjustment (Funk et al., 2011). The target estimand of the exploratory impact analysis is an estimate of the sample average treatment effect of treatment on the treated. In other words, the goal is to obtain an estimate of the average effect of Eedi on teacher homework-related workloads for those teachers that adhere to the treatment. It has to be said, however, that the extent to which any estimate produced through this analysis can be considered unbiased is likely to be limited by the circumstances and the data available. The weight used in inverse probability weighting is a composite weight constructed to achieve two objectives: 1) to return the profile of schools participating in teacher surveys at Waves 1-3, to the profile of all schools at randomisation based on observed covariates drawn from the school census;

and 2) a weight derived from propensity scores that seek to reweight the sample such that it is balanced on individual and school-level covariates across those teachers that do and do not adhere to treatment. The resulting composite weight is used to weight the analysis in which a regression model is estimated. The regression model examines the effects of adherence on workload adjusting for the same set of covariates used to model the weights (the so called doubly robust approach). The procedure is described as being doubly robust because estimates obtained from it remain consistent if either the equation used to generate the propensity scores is mis-specified or the regression equation itself is mis-specified (Funk et al., 2011). If both are mis-specified then treatment effects will be biased.

A range of further analyses, in addition to that to be performed on the teacher survey data, are proposed. Administrative data from the Eedi software management system are linked to pupils' NPD records, specifically an indicator revealing whether pupils have qualified for free school meals in the last six years (ever-FSM) and their KS2 mathematics attainment. A range of dependent variables are derived from this administrative data. These include:

- The number of quizzes each pupil in the intervention schools were set by their teacher over the two year period of the trial
- The number of quizzes each pupil in the intervention schools commenced over the two year period of the trial
- The number of quizzes each pupil in the intervention schools completed over the two year period of the trial
- The average differences between the number of correctly answered questions between the two quizzes in each pair
- The number of diagnostic questions answered by pupils in the intervention schools during the Covid-19 lockdown (from the 21st of March 2020)

Table 2: Design – exploratory impact analysis based on teacher survey data

Design		Quasi-experimental: Inverse probability weighted regression adjustment
Unit of analysis **(school, pupils)**		Teacher
Number of Units to be included in analysis **(Intervention, Comparison)**		824 teachers in total; 251 teachers report using the platform
Outcome	variable	Teacher workload in hours per week
	measure (instrument, scale, source)	Teacher questionnaire – self reported hours
Baseline for outcome	**variable**	Average teacher workload per week aggregated to the school level
	measure (instrument, scale, source)	Baselines teacher questionnaire

Participants

For the exploratory impact analysis the target study participants are teachers teaching mathematics to the focal cohorts of pupils in Years 10 (school year 2018/19) and 11 (school year 2019/20) in 158 English secondary schools recruited to the study. From this initial sample of schools, 149 distributed questionnaires to teachers. Within this sample of schools and potential pool of respondent teachers, teachers taking part in the survey provided a measure of their self-reported mathematics-related homework workload, measured in hours-per-week, for a specified reference week. Teachers in both intervention and control schools were also asked whether they were using the platform at the time of the survey. Thus reported teacher homework-related workload among the achieved sample can be linked to whether teachers reported using the platform.

Unfortunately, researchers were unable to obtain a sampling frame for teachers in range of the intervention and in schools assigned to intervention and control conditions. As a result reliance was placed on schools participating in the study distributing questionnaires to teachers. This was achieved through schools emailing a link to the questionnaire at baseline, that is prior to or just after randomisation, at December 2018, March 2019 and March 2020.

For the further exploratory analysis described in this protocol, study participants are children studying mathematics GCSE in Years 10 and 11 over the period September 2018 to 20th March 2020 (when schools were shut in response to the Covid-19 pandemic) in intervention schools. This further analysis looks at various measures of pupil adherence in relation to the intervention.

Sample derivation and sample size calculations

Table 4 provides an estimate of the minimum detectable effect size for the exploratory impact analysis conducted using teacher survey data. In order to derive an MDES we have had to make some decisions about how the analytical sample is to be constructed from the survey data and draw on existing evidence of the likely correlations between pre and post-test outcome measures as well as values for the ICCs. Because we have so little evidence from previous teacher level studies conducted within English schools, values used in these calculations are to some extent are based on judgements about what is plausible.

The assumptions made is our calculations are as follows:

- Two sided test of statistical significance
- Long run error rates set to 5 per cent (Type I) and 20 per cent (Type II) respectively
- Correlation between baseline workload covariate at the school level and individual teacher level work load at follow-up 0.4
- ICC, set to 0.2 which is conservative and consistent with values typically used in the planning of EEF trials
- Proportion of teachers taking-up Eedi, and thus in the ‘intervention group’, in the achieved teacher survey sample is 28 per cent[3]

[3] This is a weighted average of the proportions at each Wave that say they are using the Eedi platform and replaces the proportion treated input into the minimum detectable effect calculation.

Table 3: Longitudinal pattern of response to teacher survey by intervention and control schools

	Intervention	Control	Total
Total number of teachers supplying observations Waves 1-3	**323**	**501**	**824**
Number of schools	70	79	149
One observation	224	292	516
Two observations	82	153	235
Three observations	17	56	73
Total teacher observations (teachers by number of observations supplied)	**439**	**766**	**1205**

The sample sizes are those achieved, and are based on the anticipated approach to deriving an analytical data set from the survey returns. In order to simplify some aspects of the analysis we propose to structure the survey data for analysis by taking the following steps:

- First each response received in Waves 1-3 (excluding baseline) will be merged into a single file
- Responses will be linked across individuals so that responses at each Wave 1-3 can be related to an individual teacher
- An individual teacher record is then created: for those teachers that supply only one observation across Waves 1-3 their single observation is entered into the data file; for those that supply two or three observations on teacher workload we take an average of these. This step is taken to remove clustering of repeated observations within teachers. This simplifies the data structure whilst making use of all survey returns received from individual teachers and is in effect an analysis of cluster means, where the cluster, or lowest level in the data is the teacher. A variable capturing the number of observation supplied by each teacher is entered into all statistical models discussed below.

These steps lead to a data file containing 824 teachers across 149 schools. MDES calculations are performed in PowerUp. The assume impact regression is an individual teacher level equation with full treatment by covariate interactions and cluster robust standard errors.

Please note that these MDES estimates are indicative rather than predictive and therefore caution is advised in how they are interpreted. Specifically, the MDES estimates show the sensitivity of the proposed analysis with the specified sample size (schools & teachers). The sample sizes relate to a different population to the original design (teachers instead of pupils) with notable attrition in the context of the Covid19 pandemic (i.e. not random). These indicative MDES estimates are provided here to provide some consistency with the approach taken in the first Eedi trial SAP but should not be regarded as of equivalent validity.

Table 4: Sample size calculations based on exploratory impact analysis using teacher survey data

		Teacher
Minimum Detectable Effect Size (MDES)		0.20
Pre-test/ post-test correlations	level 1 (teacher)	n/a
	level 2 (school)	0.40
Intracluster correlations (ICCs)	level 2 (school)	0.20
Alpha		0.05
Power		0.8
One-sided or two-sided?		2
Average cluster size		5.5
Number of schools	Intervention	70
	comparison	79
	Total	149
Number of Teachers	Intervention	323
	comparison	501
	total	824

Outcome measures and covariates

The data used in the exploratory impact analysis come from a survey of teachers and from aggregate published school statistics linked to the survey data responses.

The design of the teacher survey, the questionnaire and its administration are described in the orginal trial protocol (Seymour & Morris, 2018). In brief, an online questionnaire was administered to teachers in control and intervention schools prior to commencement of the intervention (June/July 2018) and again post intervention in December 2018, March 2019 and March 2020. We refer to the first of these measurement occasions as the baseline, and subsequent measurement occasions as Waves 1 through to 3. Schools randomised to both intervention and control were asked to email the questionnaire to teachers that met the inclusion criteria for this study; that is all teachers with responsibility for teaching GCSE mathematics to the focal cohort of pupils, in either intervention or control schools, at the point in time the survey was administered. To give a sense of the response rate; of the 1663 email addresses available at the baseline, 87 contained errors and were undeliverable; and 137 teachers that received an email indicated that the survey was not relevant to them. In total 634 useable questionnaires were returned, 307 from intervention group teachers and 327 from control group teachers, representing a total response rate of 38 per cent (in 149 schools). But because no accurate sampling frame was available we do not know how many available email addresses were related to a teacher that met the inclusion criteria for the study, nor generally the total number of teachers that met criteria for inclusion in the sample population. Thus no reliable or accurate response rate can be calculated at the teacher level, and this is a major limitation of the proposed analysis.

Questionnaires at each measurement occasion asked teachers teaching mathematics to the focal cohort of 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 outcome for the exploratory impact analysis – homework related 'teacher workload' – is constructed from the data by adding up the amount of time spent on each of the tasks set out above to provide a measure of total workload in a reference week measured in minutes.

As mentioned, the survey questionnaire was administered at baseline to Year 10 and 11 teachers during the summer of 2018. From this survey it was hoped that a pre-intervention workload measure might be obtained for individual teachers. Because so few teachers supplied an observation on workload at Waves 1-3 and at the baseline, such a covariate were it to be derived, would have unacceptably high levels of missingness. As a result the baseline observations of work load among Year 10 and 11 teachers are aggregated to the school level and used as a school-level workload covariate.

In addition to the outcome measure, a number of teacher-level covariates available for statistical adjustment are collected through the survey data. The following covariates are available, these are measures that we can be sure are exogenous to the intervention but are likely to be associated with take-up of the Eedi platform and correlated with workload:

- Total years' teaching experience
- Years worked in the current school
- Years worked in current role
- Role in school (classroom teacher, department head, etc)
- Whether a Newly Qualified Teacher
- Whether a maths specialist teacher
- Working hours: full time / part-time
- Gender
- What years they teach maths to (7 through 11)
- Past use of similar platforms

Further to these variables school-level-covariates from the school Census and NPD are used in these analyses. Again, it is hypothesised that these covariates are associated with teacher workload and whether a teacher reports using the Eedi platform:

- % of students achieving a level 4 (or equivalent) or higher score in mathematics GCSE in 2018, 2017 and 2016
- % of students qualifying from free school meals in 2018
- School size
- Pupil/teacher ratio
- School type

Estimation of treatment effects

As previously described, the study was initially designed as a cluster randomised controlled trial. This means that over repeated identical experiments, in expectations, teacher samples by intervention and control groups, where no loss to follow-up nor initial non-response was encountered, will be in perfect balance in terms of both observed and unobserved covariates. As a result, in any one replication of the experiment an unbiased estimate of the effect of Eedi on teacher work load can be obtained and a measure of uncertainty around that estimate derived in a relatively straightforward manner.

Unfortunately, as the numbers displayed in Table 1 reveal, the absolute size of the achieved samples by intervention and control group are quite unequal suggesting that balance across covariates by intervention and controls groups, in expectations, is unlikely. Furthermore, we also see that among intervention group teachers an appreciable proportion did not use the Eedi platform. In other words, they did not adhere to the intervention protocol. The reasons for this non-adherence will be discussed in the implementation and process evaluation that accompanies the analysis discussd here. As a result, any simple experimental comparison of average levels of workload by teachers in intervention and control group schools using the survey data is likely to be biased on two counts: 1) due to survey non-response; and 2) as a result of non-adherence. In effect, the original design of this trial assumed the estimand for the analysis of the teacher survey data was the average causal effect of intention to treat. Here we switch attention to the average effect of treatment on the treated, which becomes the target estimand and which we address directly, rather than through estimating complier average causal effects.

In this section we describe a exploratory impact analysis that seeks to address these problems; though one as previously noted with some significant limitations. A series of additional regression models and sensitivity checks that aim to aid in the interpretation of results and which support the exploratory impact analysis are described. Discussion considers (a) how weights for use in the analysis are to be derived and (b) the proposed means of obtaining effect estimates via outcome regression adjustment and the conversion of estimated effects to standardised effect sizes.

Deriving weights for use in regression adjustment for treatment effects

As mentioned above, the exploratory impact analysis aims to obtain a sample estimate of the average effect of treatment on the treated. This means the average effect on teacher workload for teachers that report using the Eedi platform. This sample estimate is to be obtained from an inverse probability weighted regression model using the treatment effects suite of commands in STATA v16. .

The weights for use in the analysis are a combination of two further weights. The first corrects for non-response into the sample at the school level (we cannot derived non-response weights at the individual teacher level which would be preferred). This weight effectively weights-up responses from teachers in schools that have characteristics similar to those schools from which no teacher responses were obtained at Waves 1-3, and weights-down observations from teachers in schools over-represented in the achieved sample. The second weight is a set of propensity scores or the predicted probability that a sampled teacher reports using the platform. From these predicted probabilities a weight is obtained that weights-up the influence of teachers that did not use the platform but who have a high predicted probability and weights-down the influence of teachers not using the platform that have low predicted probability.

For the average effect of treatment on the treated, Morgan & Winship (2015) suggest the following equations for such composite weights:

For $T_{ij} = 1$:

$$\omega_{ij} = r_i \times 1$$

For $T_{ij} = 0$:

$$\omega_{ij} = r_i \times \frac{p_{ij}}{1 - p_{ij}}$$

Where $\omega_{ij}$is the weight to be used in the regression analysis described in the following section. T_{ij} is set to one if teacher j in school i reports using the Eedi platform at the time of the survey, zero otherwise. The weight r_i is the inverse of the response probability for school i. This captures the probability that at least one observation on teacher workload is received from the school. It is a obtained from logistic regression model estimated on the sample of 158 schools initially randomised to intervention and control conditions as part of this study, with a dependent variable coded to one if at least one observation on teacher workload is obtained from the school, zero otherwise. The model will include as covariates the five school level factors mentioned in the previous section.

The weight p_{ij} is the probability that teacher j in school i reports using the Eedi platform and is obtained from an individual teacher level logistic regression model where the dependent variable is equal to one if teacher j reports using the platform, zero otherwise, with both individual level and school level factors discussed in the previous section included as covariates. In other words $p_{ij} = P(T = 1|X^T)$ obtained from a logistic regression model[4].

The logistic regression equation used to derive propensity scores includes a range of covariates from the teacher survey and school census. It should be noted, however, that the teacher questionnaires were designed in the context of what was originally an experimental study and not with the modelling of selection into intervention in mind. Therefore the available covariates that can be derived from questionnaire data are likely to be less than optimal for the purposes of modelling selection into the intervention. Given that we have only anecdotal evidence of the factors that might be driving take-up of the Eedi platform and school-level non-response, we start from the position of including all the available teacher-level covariates mentioned above in our analysis[5]. In terms of covariates extracted from the school census, these attempt to capture both the prior performance of the school in national mathematics examinations, the size and type of school and the level of deprivation among students. We hypothesise that prior attainment is likely to be associated with teacher workload and survey response, as might levels of deprivation.

Estimation of propensity scores

A first run of the propensity score equation will include the full set of individual teacher and school level covariates as main effects only (X^T). The balance achieved will be assessed by weighting the sample by the propensity score from the initial run and calculating the

[4] The logistic regression models will be estimated using the STATA v16 command 'logistic', with cluster robust standard errors; although for the purpose of deriving propensity scores derivation of cluster/robust standard errors is not strictly necessary

[5] Note Austin & Stuart (2015) suggest that it is not appropriate to use statistical tests in the analytic sample in order to identify covariates for inclusion in the estimation of propensity scores. This is due to the possibility of low statistical power in attempting to identify important covariates and that further the objective is to achieve balance in the sample not to make inferences to the population.

standardised mean differences by intervention and control group for each covariate in the model, and taking the average of these across the covariates as well as the difference in scaled standard deviations for each continuous covariate (Morgan & Winship, 2015). A visual inspection of the distribution of estimated propensity scores by intervention and control groups will be undertaken in order to assess common support and a box plot of the scores presented in order to examine any extreme scores. Various changes to the propensity score equation will be considered on subsequent runs of the model. These changes will include interaction terms and higher order terms for continuous covariates with the possibility of extending this to cubic spline functions (Austin & Stuart, 2015; Harrell, 2015). Adjustments to the specifications will be considered in sequence and after each change to the model the standardised mean differences by intervention and control group for each covariate in the model will be calculated again as well as their average across the covariates and differences in scaled standard deviations. Further visual and graphical assessments of results will also be undertaken. These metrics will be assessed iteratively until a specification for the propensity score equation is found that minimises imbalances in the observed covariates and maximises common support. The final chosen propensity score model will reported in an Annex to the final report.

ANALYTICAL MODEL FOR THE EXPLORATORY IMPACT ANALYSIS

For the inverse probability weighted regression model, the following equation is proposed:

$$Y_{ij} = \beta_0 + \beta_1 T_{ij} + \beta_2^T X^T + \beta_3^T T_{ij} X^T + \varepsilon_{ij} \ldots\ldots[1]$$

Where Y_{ij} is average hours per week for teacher j in school i; T_{ij} is set to equal one if teacher j in school i reports using the platform school, zero otherwise; and X^T are teacher and school level covariates as described above. The specification allows for the fact the sample is unbalanced in the Xs as a result of selection into treatment in that a full set of treatment by covariate interactions are specified. Here β_2^T and β_3^T are vectors of coefficients. If the Xs are centred at the means for those who use the platform (the treated sample), β_1 is the sample estimate of the average treatment effect on the treated (Schafer & Kang, 2008). If the Xs remain uncentred then the sample estimate of the average effect of treatment on the treated is the average in the predicted margins across increments in T_{ij}. Estimates for the parameters in Equation [1] are obtained where the sample is weighted by ω_{ij} as described in the previous section.

The analysis rests on a number of assumptions: 1) that the Xs include all potential confounders; 2) that the relationship between the potential outcomes and covariates in models for the propensity scores and non-response weights are correctly specified in addition to including all confounders; 3) that each teacher has a non-zero probability of Eedi platform use; and 4) that the distribution of propensity scores in intervention and control groups have sufficient overlap.

All standard errors and confidence intervals will be derived using robust methods and take account of clustering of teacher observations by schools. The parameter for the effect size is as follows, assuming covariates centred at the mean value for the treated portion of the sample:

$$ES = \frac{\beta_1}{S}$$

Where β_1 is the coefficient on the treatment indicator in equation [1] and S is the total, unconditional, pooled within group standard deviation. The standard error for the sample estimate of the effect size will be obtained using bootstrap procedures based on a 500 replications. Uncertainty in the effect size estimate will be reported in the form of a frequentist 95 per cent confidence interval.

SENSITIVITY AND OTHER SUBSIDIARY ANALYSIS

In order to provide adequate context for the impact analysis and to aid in the interpretation of results a range of analysis describing the sample is planned and will be presented in the report of findings prior to the estimation of treatment effects. In addition, further regression models, weighted analysis and additional sensitivity checks are proposed.

Descriptive analysis

The following descriptive analysis of the **raw sample of teachers** for analysis (that is the unweighted sample) will be performed and presented in tabular form in order to provide context for the estimation of the exploratory treatment effects (see section on sample size calculations for the derivation of the sample for analysis):

- For the outcome and continuous covariates the mean and standard deviations by intervention and control group
- For categorical covariates, proportions by intervention and control group
- Standardised mean differences between intervention and control groups for all covariates

Following this, for the **sample weighted by propensity score** from the best performing specification of the propensity score equation:

- For the outcome and continuous covariates the mean and standard deviations by intervention and control group
- For categorical covariates, proportions by intervention and control group
- Standardised mean differences between intervention and control groups for all covariates

Specification checks - the regression analysis

After the presentation of the estimated effects and in order to inform interpretation and judgements regarding the reliability of results, we propose to present results from a series of further regression models. Estimated effects from these models will be compared to results from the main analysis in order to determine the sensitivity of the results to differences in model specification. These additional regressions will be:

- Simple bivariate regression of teacher workload on the treatment indicator
- Simple bivariate regression as above but estimated using weights ω_{ij}
- Regression model will full set of covariates but direct effects only; that is with no treatment by covariate interactions
- Regression model with full set covariates and all treatment by covariate interactions but not using weights ω_{ij}

Specification checks - the propensity scores

To assess the potential consequences of extreme propensity scores and the extent of common support in our sample on the results of the exploratory impact analysis, we plan to re-run the analysis in two ways:

- Generate three sets of trimmed weights, trimming ω_{ij} at the 99th and 1st percentiles, the 95th and 5th percentiles and the 90th and 10th percentiles. Equation [1] would be re-estimated using each set of trimmed weights and the effects of extreme weights on sample estimates assessed
- Re-estimate equation 1 using inverse probability weights but on a sample where common support has been imposed – it is likely that cases in the intervention sample whose propensity scores exceeds the maximum in the control group will be pruned from the sample as well as those cases in the control group whose propensity score falls below the minimum in the intervention sample (Caliendo & Kopeinig, 2008).

CHECKS FOR THE POSSIBLE EFFECTS OF HIDDEN CONFOUNDERS

One of the central assumptions that underpins the exploratory impact analysis is that we have a set of covariates that enables us to estimate the correct propensity scores. Thus the extent to which the proposed analysis produces unbiased estimates can be undermined, among other things, by the existence of some unmeasured or hidden confounder U. Rosenbaum (1986, 2017) provides techniques to assess the extent to which a hidden confounder may, if it could be incorporated into the analysis, change the results. West et al. (2014) provides a summary of practical guidance in terms of conducting such an analysis drawing on earlier work from Hong (2004).

We will assume that a hidden covariate is unaccounted for in our analysis with a standardised mean difference between intervention and control groups in the unweighted sample equal to the largest difference observed among the measured covariates (Hong, 2004). According to Rosenbaum, (1986) this standardised mean difference should be re-scaled for use in sensitivity analysis by dividing by $\sqrt{2}$. We also need to make an assumption about $\gamma,$ the extent to which U is correlated with the outcome Y given the propensity score. Hong (2004) suggests a value for γ be set initially to equal the largest correlation with the outcome among the measured covariates. Once assumptions have been quantified in this manner, the estimated effect size from the main analysis can be adjusted such that the estimate is based on equating all covariates plus U using the following equation (West et al., 2014):

$$ES^* = \widehat{ES} - \gamma\left(\frac{d_U}{\sqrt{2}}\right)$$

Where d_U is the assumed standardised mean difference for the hidden covariate U in the unweighted sample and $\widehat{ES}$ the estimated effect size from the exploratory impact analysis. Statistical significance can be evaluated by dividing ES^* by the standard error of the effect size from the impact analysis. The assessment will be conducted where results from the impact analysis reach standard thresholds for statistical significance; various plausible scenarios regarding U and γ will be examined.

Further exploratory analyses

We propose several sets of further analyses based on the data collected on pupils in intervention schools. The analyses are observational and exploratory in nature.

Analysis of Eedi administrative data over the course of the trial

The analysis will be carried out using administrative data made available to the research team as part of the process evaluation. Although descriptive analyses of these data were proposed in the February 2020 SAP, what is discussed here is considerably more complex. We will delve deeper into the data to understand how pupils' characteristics are linked to key usage indicators. The models will adjust for the lack of take up of the intervention across the study sample (i.e. a large preponderance of zero in the dependent variables) and we also envisage estimating a wider range of models with different dependent variables to that originally proposed, where the focus was only on the number of quizzes started.

The dataset to be used in these additional analyses was produced for this evaluation by Eedi through extracting and aggregating information from their operational database. This data extract includes information on the pupils that used the Eedi platform over the two years of the trial, including the school and class they were in[6]. Usage information is recorded as follows (we propose to use these as outcome variables to be explained in our analysis): the number of assignments set, started and completed, along with other information such as whether feedback was provided and whether it was read; the number of follow-up questions set and answered; and, the number of mistake resolved. Information on the interaction with parents is also available.

To carry out the analysis, in addition to the dependent/outcome variables mentioned, we will link the data with an extract from National Pupil Database containing the following variables which will be used as covariates in our proposed regression models (discussed further below):

- Sex (pupil level)
- Ever-FSM (pupil level - from NPD)
- Month of birth (pupil level)
- Points score at KS2 mathematics (pupil level – from NPD)[7]

We will use regression analysis to estimate the impact of the explanatory variables on the outcomes. Particularly, we will make use of two-part models that enables us to account for the large number of zeros present in the data, which stem in turn from the relatively low take-up of the platform. For example, approximately 28 per cent of pupils in intervention schools were not set any assignments; and 45 per cent of all pupils did not complete any assignments (this is reduced to 24 per cent if we count only those who were set an assignment). A two-part model is a model which accounts for dependent variables that are a combination of continuous response and a mass of observations at zero (Belotti, Deb, Manning, & Norton, 2015; Deb, Norton, & Manning, 2017). The expected mean can be written as the product of the expectations from the two parts of the model (Belotti et al., 2015):

[6] These data permit the examination of class to class moves that occur during GCSE, a subject about which there is currently limited evidence. Evidence on the extent of class to class moves is likely to be of interest to those designing cluster randomised trials in education. For this reason, EEF are currently in discussions around the potential for further analyses addressing this topic.

[7] We will record the proportion of records linked successfully to NPD.

$$E(y|x) = \Pr(y > 0|x) \times E(y|y > 0, x)$$

For the second part, a regression model for a continuous outcome can be used, such as OLS, poisson or negative binomial, while the first part specified as a binary response can be estimated using a logit or probit link function. The error terms in both parts of the model need not be independent for consistent estimates of parameters, and the covariates x are the same in both models. For example, we will use this set up to model outcomes such as number of quizzes a pupil is set, where the probability that the number of quizzes is at least one is modelled using a logistic link function and a count of the number of quizzes set conditional on at least one assignment being set is modelled assuming a negative binomial distribution.

Given that pupils are nested in schools, the model will either also include school fixed effects or we will specify cluster robust standard errors (school random effects). In essence, the two-part model generates two sets of regression coefficients, one for each part of the model. However, it allows for marginal effects and their standard errors to be computed taking into account both parts.

The analysis will be implemented using the Stata program ‘`twopm`’.

Analysis of test pairs

The Eedi platform incorporates functionality that allows for two quizzes to be set on the same area of subject knowledge consecutively with remedial actions between quizzes. The first quiz is set right after teaching the lesson (this is named test A) and the second test is set 3 weeks later (this is named test B). In the weeks between the two tests pupils can receive feedback, homework or parents can be notified.

The aim of this analysis is to examine whether the number of correctly answered questions increases in test B compared to test A. In essence we will be examining the change in A and B test scores and the associations between this change and the covariates mentioned above. From this we might infer how far teaching responds to mis-understandings in the first test (test A) and thereby an improvement in pupils performance is observed (in test B).

The analysis will use detailed platform activity data. The current dataset provided by the developer (and presented in the description of the previous analysis) includes data that was aggregated at pupil level and does not contain information on performance. The developer has agreed to provide us with granular data that will include detailed information on each test completed by each pupil. This will allow us to derive variables that will quantify:

- the number of correctly answered questions per quiz;
- the change in the number of correctly answered questions between quizzes;
- the amount and type of interaction with the platform between each pair of quizzes.

We will include the following covariates in a regression model that will have the difference in Test A and B scores as the dependent variable:

- Sex (pupil level)
- Ever-FSM (pupil level - from NPD)
- Month of birth (pupil level)
- Points score at KS2 mathematics (pupil level – from NPD)

- School type (School level)
- School size Year 9 (School level)
- School prior attainment rates at GCSE mathematics (School level)
- School FSM rates (School level)
- School EAL rates (School level)
- School SEND rates (School level)

As a first step we will carry out descriptive analyses to assess how far students' observed characteristics vary by whether they complete only one test of an available pair of tests, thus determining the extent to which results of the analysis are affected by selection/collider bias, at least in terms of observables.

Second, we will use regression analysis to estimate the impact of the independent variables and covariates on the outcome. The unit of analysis will be the test pair. Test pairs are clustered in pupils, who are clustered in classes, clustered in schools. A 3-level hierarchical linear model should be used with fixed effects for school.

Analysis of Eedi usage during the Covid-19 Lockdown

There are approximately 350 pupils from the schools included in the intervention group who have continued to use the Eedi platform during the Covid-19 lockdown (after March the 21st). On average, each pupils answered 78 questions (with an SD of 79).

We propose an exploratory analysis to model the take up of the Eedi platform across schools and pupils during the lockdown – this is both those who have newly taken-up the platform and existing users who continue to use it. The analysis aims to increase our understanding of the socio-demographic factors that can influence the use of an e-learning platform in the unusual situation induced by the lockdown. The analysis will contribute to the question of whether remote learning is linked to socio-demographics.

The developer will provide us with a data extract that will include the number of questions set and answered. We will also require information on whether the questions were answered correctly.

We will use the following explanatory variables:

- Sex (pupil level)
- Ever-FSM (pupil level - from NPD)
- Month of birth (pupil level)
- Points score at KS2 mathematics (pupil level – from NPD)[8]
- School prior attainment rates at GCSE mathematics
- School FSM rates
- School EAL rates
- School SEND rates
- Deprivation

[8] We will record the proportion of records linked successfully to NPD.

As a first step in the analysis we will assess the distribution of quiz-taking by schools. Following this we will carry out descriptive analyses to measure the association between usage of the platform during lockdown and each of the covariates named above.
This will be followed, given the dependent variables are counts, by the estimation multi-level poisson/negative binomial models to estimate the simultaneous impact of the covariates on measures of usage.

This analysis will be particularly pertinent given the priority of online provision of education in response to the Covid-19 pandemic. It will reveal patterns in voluntary engagement with online provision and how this varies across and within schools. The limitation of the analyses, is that the schools in our sample had already been offered access to the Eedi platform prior to the lockdown. They are not schools that have proactively sought access to the platform in response to the lock down.

References

Austin, P. C., & Stuart, E. A. (2015). Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Statistics in Medicine*, *34*(28), 3661–3679. https://doi.org/10.1002/sim.6607

Belotti, F., Deb, P., Manning, W. G., & Norton, E. C. (2015). twopm: Two-part models. *The Stata Journal*, *15*(1), 3–20.

Caliendo, M., & Kopeinig, S. (2008). SOME PRACTICAL GUIDANCE FOR THE IMPLEMENTATION OF PROPENSITY SCORE MATCHING. *Journal of Economic Surveys*, *22*(1), 31–72. https://doi.org/10.1111/j.1467-6419.2007.00527.x

Deb, P., Norton, E., & Manning, W. (2017). *Health econometrics using STATA*. College Station, TX: Stata Press. Retrieved from https://stats.idre.ucla.edu/stata/dae/negative-binomial-regression/

Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., & Davidian, M. (2011). Doubly Robust Estimation of Causal Effects. *American Journal of Epidemiology*, *173*(7), 761–767. https://doi.org/10.1093/aje/kwq439

Harrell Jr, F. E. (2015). *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer. Retrieved from https://online.stat.psu.edu/stat501/lesson/8/8.8

Hong, G. (2004). Causal inference for multi-level observational data with application to kindergarten retention. University of Michigan.

Morgan, S. L., & Winship, C. (2015). *Counterfactuals and causal inference: Methods and principles for social research* (2nd ed.). New York, NY: Cambridge University Press.

Morris, S. P., Smith, A., & Kiss, Z. (2018). *Statistical Analysis Plan Evaluating the effectiveness of Eedi formative assessment programme (previously Diagnostic Questions) on raising attainment in mathematics at GCSE*. London, Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Projects/EEDI_SAP_2018.11.29_FINAL.pdf

Rosenbaum, P. R. (1986). Dropping out of high school in the United States: An observational study. *Journal of Educational Statistics*, *11*(3), 207–224.

Rosenbaum, P. R. (2017). *Observation and experiment*. Cambridge, Massachusetts: Harvard University Press.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Schafer, J. L., & Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods*. Schafer, Joseph L.: The Methodology Center, 204 East Calder Way, Suite 400, State College, PA, US, 16801, jls@stat.psu.edu: American Psychological Association. https://doi.org/10.1037/a0014268

Seymour, K., & Morris, S. P. (2018). *Trial Evaluation Protocol: Evaluating the effectiveness of Eedi (previously Diagnostic Questions) formative assessment programme*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/diagnostic-questions/

West, S. G., Cham, H., Thoemmes, F., Renneberg, B., Schulze, J., & Weiler, M. (2014). Propensity scores as a basis for equating groups: Basic principles and application in clinical treatment outcome research. *Journal of Consulting and Clinical Psychology*, *82*(5), 906–919. https://doi.org/10.1037/a0036387