Research protocol: Evaluating the impact of Eedi formative assessment online platform (formerly Diagnostic Questions or DQ) on attainment in mathematics at GCSE and teacher workload

Declaration of interest: the preparation of this protocol and the trial described are funded through a grant made available by the Educational Endowment Foundation

Note: the content of this protocol has been approved by the Education Endowment Foundation

Research Protocol: Evaluating the impact of Eedi formative assessment online platform (formerly Diagnostic Questions or DQ) on attainment in mathematics at GCSE and teacher workload

Abstract: This paper describes a cluster randomised controlled trial designed to test the efficacy of Eedi, a formative question setting and diagnostic digital platform, on attainment in mathematics at GCSE as well as its impact on teacher workload. The study is a pragmatic two arm trial that aims to randomise 180 English secondary schools to intervention and control (business as usual) conditions. The intervention is targeted at Year 10 pupils (aged 14-15 years) and their teachers commencing study of GCSE mathematics from September 2018 and will run for two years. The study is due to report at the end of 2021.

Key words: research protocol, formative assessment, student feedback, online question setting, mathematics, cluster randomised controlled trial

1. Background

This protocol describes a pragmatic cluster randomised controlled trial, the aim of which is to assess the efficacy of an online formative assessment and feedback program to be used to set, mark and provide feedback on mathematics homework in Years 10 and 11 of the English school system. The intervention, known as Eedi (formerly Diagnostic Questions or DQ) will be implemented in approximately 90 English secondary schools in the Autumn term 2018 and run for two years. The aim of the intervention is twofold: 1) to raise attainment in GCSE mathematics; and 2) reduce teacher workload, particularly that related to maths homework, such as setting, marking and student feedback. Attainment at GCSE in mathematics will be the primary outcome. A further 90 schools will act as controls and implement 'business as usual' strategies toward raising attainment and addressing teacher workload. The intervention will be implemented and delivered by Eedi, a commercial education technology company, in partnership with the Behavioural Insights Team (BIT), a social purpose company. The evaluation will be conducted by AlphaPlus Consultancy in partnership with Manchester Metropolitan University.

The main purpose of this evaluation is to provide evidence as to the effects of Eedi on attainment. Assessing the existing literature in seeking to understand the likely impacts of such an intervention is, however, fraught with difficulty. The terms formative assessment or feedback cover a wide variety of learning practices (Bennett, 2011). Tools used can be computer-based or involve a range of other technologies. Apart from differences in the mode through which

formative assessment can be delivered, the subject matter, extent and type of feedback, whether tools are used in the classroom or outside of schools online, the degree to which formalised measurement principles inform assessment, stage in schooling, whether assessments are adaptive or not, involvement of parents, can and do vary considerably from application to application.

What is clear is that there is no existing evidence that specifically assesses the effects of Eedi. It is an unproven technology. Furthermore, the sheer extent of heterogeneity makes drawing conclusions as to what such a varied literature is saying, and thus forming expectations about the likely efficacy of Eedi, difficult. Studies in this field of enquiry have been consistently criticised for their poor quality and few appear to be relevant to the United Kingdom (UK) or English context.

As far back as the late-1990s, reviews of the evidence reported large effect sizes associated with a wide range of formative assessment interventions, where formative assessment was defined as the gathering of information to be 'used with the intent of assisting in the learning and teaching process' (Black & Wiliam, 1998, page 29). Later studies cast doubt on such findings, particularly the magnitude of effect sizes but also the consistency of findings, definitional issues and quality of the evidence base (Bennett, 2011; Kingston & Nash, 2011; Van der Kleij, Feskens, & Eggen, 2015). For example, in their review Kingston & Nash (2011) could only find 42 'useable effect sizes' in the literature since 1988 (from 13 studies). They found that formative assessment 'can be a significant and readily achievable source of improvement [in] student learning' (Kingston &

Nash, 2011, page 33) but that effect sizes are likely to be lower than previously reported (weighted mean effect size = 0.20, 95% confidence interval 0.19-0.21) and that there was wide variability in effectiveness. Of particular relevance to this study was the authors' finding that 'computer based formative feedback interventions produced a mean effect size of .28 (95% confidence interval of .26 to .30). However, Kingston & Nash (2011) review has also been criticised on grounds of not paying sufficient attention to the quality of studies included in their review and the variety of interventions described as formative assessment – confirming again that definitional challenges and issues of study quality plague this area of research (Bennett, 2011; McMillan, Venable, & Varier, 2013; Rakoczy et al., 2018).

In short the research suggests that formative feedback shows some promise but that adherence and quality of feedback matters (Konishi, Wong, & Tao, 2018; Pinger, Rakoczy, Besser, & Klieme, 2018). It is also pointed out that teachers require appreciable levels of knowledge in order to use formative assessment effectively, requiring support and necessitating non-trivial levels of investment in professional development (Bennett, 2011).

But what specifically of computer-based formative assessment and feedback interventions? Van der Kleij et al. (2015) undertook a meta-analysis of formative feedback within a 'computer-based environment'. They note the possibility that computer-based feedback and assessment can provide students with individual assessment of test/question performance in a timely manner with scoring and assessment undertaken automatically. Thus they argue, computerized assessment can bring teaching closer to the goal of more personalised feedback

and support. The authors looked at various forms of feedback on attainment – Knowledge of Result (KR), Knowledge of Correct Response (KCR) and Elaborate Feedback (EF) compared to each other as well as no feedback. They found that:

> "more elaborate feed-back led to higher learning outcomes than simple feedback, in particular in regard to higher order learning outcomes" (Van der Kleij et al., 2015, page 505)

Their meta-analysis found an overall effect size of 0.49 for EF compared to simpler forms of feedback and no feedback (effect sizes for KR and KCR were 0.05 and 0.32 respectively).

Shute & Rahimi (2017) consider what they term computer-based assessment for learning (CBAfL), which comprises formative assessment and summative elements. The authors identify eight studies of which three relate to web-based CBAfL and are therefore relevant in the context of this present evaluation. The authors claim that computerized formative feedback provided in the classroom can aid attainment if feedback strikes the right balance between providing enough detail but not being overly-complex. Second, that use of computerized systems of formative feedback and assessment needs to be sustained over time. Looking specifically a web-based systems similar to Eedi, Shute & Rahimi (2017) argue that web-based systems play an important role in keeping students engaged in their learning and as a way that students can monitor their progress independently. Online-based systems were commended for promoting accessibility to learning and their ease of use. Analyses of data derived from computer-based learning systems can be examined to find 'hidden learning and

error patterns' (Shute & Rahimi, 2017, page 15) as well as confirm understanding that aid personalised learning and feedback.

A secondary factor motivating this study is the perceived problem of excessive teacher workload. The automated assessment, marking, diagnostic and feedback features of Eedi are hypothesised to reduce teacher workload as it relates to mathematics homework. The 2016 Teacher Workload Survey revealed that on average secondary school classroom teachers spent 8.1 hours per week marking or correcting pupils' work of around 33 hours per week spent in total on non-teaching tasks (Higton et al., 2017). This was considered to be too long by teachers responding to the survey. Over half of respondents (52%) considered workload to be a serious problem and three-quarters were dissatisfied with the hours they usually worked[1]. The Independent Teacher Workload Review Group claimed in its 2016 report that providing written feedback on pupils work had become excessively burdensome for teachers, and unnecessarily so. The report questioned whether extensive written comments on every piece of work is effective at raising outcomes in the long run and suggested that providing 'excessive' written feedback distracts teachers from more important aspects of their work. The Education Endowment Foundation's own review (Elliott et al., 2016) noted that marking is one of the main drivers behind what is seen as excessive teacher workloads. Furthermore, that there was little high quality and reliable evidence on whether marking had an appreciable role in raising attainment.

[1] It needs to be kept in mind that the response rate to this survey was 34 per cent and therefore estimates may be biased.

The evidence, though somewhat difficult to summarise, suggests that digital platforms such as Eedi, that provide continuous formative assessment and feedback as well as in-built diagnostic elements, may have the potential to raise attainment within the English context. However, it is more realistic to conclude that the benefits to both pupils and teachers have not been established and it is uncertain as to whether such platforms will contribute the raising of attainment within English schools. More, specifically, the efficacy of the Eedi platform as a technological solution consistent with formative and diagnostic elements has not be demonstrated. In short, the literature is plagued by ambiguity, definitional issues and poor quality studies. It is also plausible that a system such as Eedi might be effective at reducing teacher workload. But again the existence of such effects has not been established through high quality randomised studies.

2. Intervention

The intervention is an online question setting and diagnostic platform which takes the form of weekly 'quizzes' delivered to pupils as homework and marked automatically within the computer program.

There are a number of popular online computer programs used in the teaching of mathematics in English schools. The Behavioural Insights Team selected the Eedi platform for trial for both substantive and pragmatic reasons. Substantively, particular features of the Eedi platform, particularly its diagnostic elements, were understood to be consistent with notions of formative assessment as discussed in the literature and there was a desire to assess the

role online platforms, such as Eedi, might play in this regard. Second, pragmatically, the developers of the Eedi platform were willing to work with both the Behavioural Insights Team and the independent evaluators in bringing their platform to trial. This is clearly an important consideration given the practical requirements of subjecting an intervention such as Eedi to rigorous testing.

The Eedi program has a diagnostic component that identifies weaknesses in pupils' understanding based on their responses to questions. The implied theory is that the Eedi system offers a reliable assessment of students' weaknesses and provides the student appropriate feedback and resources in response. Teachers can also review pupils' performance in tests and adjust their teaching in response. The platform also affords a means by which parents can review their children's work stimulating their greater involvement.

The intervention will be introduced in intervention schools during the autumn term 2018 and will be available to Year 10 students and their teachers. The intervention will continue to be available to this cohort on entry to Year 11 up until the point they sit GCSE mathematics examinations in the summer of 2020. Over the period of the study, control schools will be barred from accessing the Eedi system; they can, however, access the system if they choose to leave the trial. It is important to note that the Eedi platform is widely available and was so prior to the launch of this study. Students will be invited to complete weekly tests in their own time as part of their mathematics homework and will be able to access the tests via a computer connected to the internet. As part of the

intervention, additional support will be provided to students who do not have access to the required networked computer device at home.

The intervention comprises four key elements. First, the Eedi system populates a set of weekly formative assessment quizzes for the entire school year aligned to the appropriate exam board scheme of work for the school. Quizzes contain 10 multiple choice questions with four possible responses (including the correct answer) to each question. Second, each wrong answer is designed to detect a specific misunderstanding. The system marks students' responses to the question and prompts the student to review their answers to questions and feedback is given through the system targeting specific misunderstandings. Students can be provided with targeted learning materials through the system addressing areas where they appear to have misunderstandings. Third, although the system automatically 'marks' students' attempts at quizzes, teachers have a review facility that enables them to also identify students' weaknesses and common misunderstandings and provide additional targeted feedback. Finally, parents can receive texts and emails setting out the quizzes their child has been set, whether the quizzes are completed by their child, and in general information on the topics are being covered in class. Parents can receive additional information regarding their child's performance in quizzes by logging on to the system.

In order to ensure effective delivery and implementation, a comprehensive programme of support and training will accompany the introduction of Eedi within intervention schools. This training is a feature of the trial and is not

usually available to users of the Eedi system. Each school is asked to appoint a project lead whose role is to liaise with the developers Eedi and BIT. Eedi/BIT work with the school-lead to help them in setting up the school's scheme of work on the Eedi platform. The school maths department is provided with two hours of training, at the school site, during the Summer term 2018 that covers: a) the importance of formative assessment; b) how to monitor quizzes and student performance; c) giving feedback to students and setting further quizzes; d) parental alerts; e) how to ensure all students have access to Eedi for home work regardless of their online access at home; and f) how to access Eedi's technical support function. School project leads will also receive training on troubleshooting the system and accessing online and other forms of support.

Throughout the trial, Eedi/BIT will monitor usage of the platform by schools and proactively approach schools with offers of support where patterns of online activity are suggestive of problems. Parents and pupils are offered Oxford University Press learning materials, addressing areas of misconception for pupils identified through Eedi, throughout the life of the trial. This is usually a 'paid for' service but for the purposes of this study is offered free of charge.

3. Research plan

The efficacy of the Eedi platform in raising pupil attainment and reducing teacher workloads is to be evaluated through implementing a pragmatic, two arm cluster randomised controlled trial (CRT), accompanied by a mixed method process evaluation. This protocol discusses the design of the CRT only. Researchers interested in the design and execution of the process evaluation are

referred to the policymaker/practitioner-focused protocol published by the Education Endowment Foundation on their website (Seymour & Morris, 2018).

3.1 Research questions

This efficacy study aims to address the following questions:

- What is the effect of exposure to Eedi on attainment in mathematics at GCSE?
- If such an effect is identified, does the effect vary by whether pupils have ever qualified for free school meals?
- If such an effect is identified, does the effect vary by sex? And
- What is the effect of exposure to Eedi on maths homework related workload for teachers?

3.2 Trial design

A CRT design, in which schools are the unit of randomisation, was chosen for three reasons. First, randomisation of individual pupils would be difficult practically due to the way homework is set for whole intact classes or sets of pupils and because of concerns that individual pupil attainment can in theory be affected not only by their own allocation to intervention and control but also the allocation status of other members of their social network and class within the school setting. The consequences of this are that individual pupils are not statistically independent of one another in potentially quite complex ways. This dependence if not recognised in the trial design and analysis of the resulting data could lead to potential biases (Bloom, Bos, & Lee, 1999; Raudenbush, 2008). Randomising at the school level can remove the effects of such biases. Second,

randomising classes would also not be practical given that teachers teach more than one class. Moreover, there would be a risk of treatment diffusion between control and intervention classes taking place within the schools regardless of whether teachers themselves or classes were randomised to intervention / control. This would occur where control classes/teachers adopted the intervention on the basis of learning about it from their colleagues assigned to the intervention. Given the nature of the intervention and the capacity to share passwords and login details within schools this would be difficult to prevent. For these reasons and despite the loss in statistical efficiency resulting from the clustered nature of the resulting data, a CRT design was chosen.

The trial design involves the recruitment of up to 180 schools to trial by the developers (Eedi/BIT). Schools were identified from records held by the Education Endowment Foundation, AQA and EdExcel exam boards and through the developers' own databases. Schools were approached by the developers in order to judge their willingness to take part in the study. Once schools signalled their interest in participating, Eedi undertook an analysis of their existing data bases to examine the extent to which pupils and teachers at 'interested' schools were already using the Eedi system. As noted above, Eedi was widely available and extensively used across England prior to this study. The study team wanted to identify schools that had low existing use of the platform such that it had not become integrated into the schools teaching effort for Year 10/11 pupils studying for GCSE. In conjunction with the developers, an upper threshold for low existing use was determined, where only schools with 30 or fewer existing accounts across the entire school were considered for inclusion in the trial.

Below this threshold, judgement as to the extent of the existing usage of Eedi was made on a case by case basis, where the absolute size of the school was taken into account before schools were deemed to be low existing users. Furthermore, the extent of activity on existing accounts was also taken into account, such that accounts were assessed as to whether they were effectively 'dormant'. Furthermore, schools had to agree on entering the trial that if they were subsequently allocated to control conditions all existing accounts on Eedi would be suspended and no new accounts could be created during the lifetime of the study. Access would be restored if the school chose to leave the study.

The schools signal their willingness to be a part of the trial through signing a memorandum of understanding (MoU) with the developers and evaluators. Schools could not participate in the trial unless they provided a signed MoU. The MoU states clearly the obligations schools in both intervention and control groups were required to meet should they wish to continue in the study. The Schools are asked to identify the Unique Pupil Number (UPN) for each student in range of the trial prior to randomisation and whose parents have not opted to remove their child from the trial. Parents have an initial two-week period in which they can withdraw their child and are informed of how they can remove their child from the trial at any point during the study beyond this period. As recruitment takes place during the Spring and Summer terms 2018, these pupils will be in Year 9 and expected to enter Year 10 from September 2018. Once a school has agreed to take part in the trial, the school Unique Reference Number (URN), region in which the school is located, UPN for each pupil within the school in range of the trial (who has not been withdrawn from the study) along

with details of which set for mathematics each pupil is in, their age, sex and free school meal status is passed to Manchester Metropolitan University, who in turn assign schools at random to intervention and control groups on a 1:1 basis.

Randomisation will be conducted in batches. This is due to the considerable training effort that is required in order to train teachers in intervention schools. The length of time required to identify, assess and recruit 180 schools is considerable and all schools in the intervention were required to be ready to commence use of the intervention by September 2018, and therefore have received training. For this reason, the developers could not delay training until all schools had been recruited and randomised. Randomising schools in batches was judged to be the most effective means of addressing this problem, as it enabled the evaluators to release details of schools assigned to the intervention in waves so that training might commence earlier than otherwise.

One limitation of the study design relates to the timing of teacher surveys. As is discussed further below, a secondary outcome for this study is homework related maths workload for teachers measured in hours/minutes per week, for a given reference week. Measures of teacher workload are to be derived from teacher surveys administered at baseline and at three further points in time subsequent to the commencement of the intervention. Ideally, a survey of Year 10 and 11 maths teachers would have been administered prior to randomisation, from which a measure of pre-intervention workload could be derived. Thus, teacher responses to the questionnaire would be unaffected by knowledge of whether their school had been assigned to intervention or control conditions.

For practical reasons relating to delays in development of the questionnaire and the prolonged period over which schools were recruited to the trial this proved not to be possible. A baseline teacher survey is instead to be conducted at the end of the Summer term 2018; prior to the intervention commencing but after randomisation had been carried out.

4. Outcome measures and instruments

The primary outcome measure for this study is pupil attainment in mathematics at GCSE. The Education Endowment Foundation, this study's funders, was establish with the objective of tackling under-achievement at GCSE in English and mathematics among disadvantaged pupils (Education Endowment Foundation, 2016). It is the practice of the Education Endowment Foundation (EEF) that when attainment is measured as an outcome variable, it should be measured using a national standardised assessment in the UK. In many cases, this will be national curriculum tests (NCTs) for primary school pupils, or General Certificates in Education (GCSEs) for secondary school students.

Adopting attainment at GCSE as the primary outcomes in studies such as that discussed here has a number of advantages. First, considerable resources are devoted to the writing and validation of GCSE questions. Second, the costs of collecting pupil level GCSE results are low compared to administering standardised tests of attainment, given that results are extracted directly from National Pupil Database. Third, unlike administering separate standardised assessments of mathematics, using GCSE attainment 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 therefore students' future opportunities. GCSEs and their 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. In this sense, the focus on GCSE attainment as a primary outcome is justified.

On the other hand, as Baird, Ahmed, Hopfenbeck, Brown, & Elliott (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); although also noting that such concerns 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.

Once a school has agreed to take part in the trial and signed the MoU, parents of pupils in range of the trial were informed about the study and given the chance

to withdraw their child there and then, or at any point in the future. The information provided to parents as part of this process signalled the intention of the study team to link pupil-level records from the trial to the National Pupil Data at the end of the study. The primary outcome measure will be obtained from pupil-level records contained on the NPD for those pupils enumerated as part of the trial and whose parents had not withdrawn them from the trial. A measure of GCSE attainment in mathematics – specifically points score in mathematics EBacc pillar – will be obtained for each pupil. The pupil level records extracted from the NPD will contain not only GCSE points score in mathematics but also each pupils' mathematics score at Key Stage 2. This later measure will act as a measure of prior attainment for each pupil. Previous analysis conducted for the Education Endowment Foundation has shown that test scores at Key Stage 2 are highly correlated with attainment at Key Stage 4 (GCSE) (Education Endowment Foundation, 2013).

The secondary outcome measure is teacher homework related workload for mathematics. The measure is obtained from responses to surveys of teachers delivered online, direct to teachers, for whom email addresses are obtained at the time schools agreed to take part in the trial and signed an MoU. Teacher surveys are conducted prior to the commencement of the intervention (at baseline) and then at three subsequent time points. At the baseline survey, teachers in both Years 10 and 11 will be asked to provide an estimate of their weekly mathematics homework-related workload, in hours/minutes per week. Year 10 teachers are surveyed subsequently at December 2018 and March 2019. And Year 11 teachers surveyed at March 2020. At each survey occasion teachers

are asked to provide an estimate of the total amount of time they have spent in a reference week (in hours and minutes): a) preparing maths homework; b) setting maths homework; c) marking maths homework; d) recording, chasing and analysing maths homework data; e) giving verbal (i.e. spoken) feedback to students based on their maths homework; f) planning maths lessons; and g) communicating with parents/carers regarding maths homework. From survey responses an estimate of total mathematics homework-related workload per week, per teacher will be computed in hours/minutes.

5. Sample

Schools were recruited by the developers based on school records held on the Eedi data base, EDUCATION ENDOWMENT FOUNDATION data bases and records supplied by AQA and EdExcel exam boards. Schools expressing interest in the study were assessed based on Eedi's administrative records as to existing use of the platform within the school, with only those schools in which usage was deemed low to minimal eligible for entry to the trial. Schools that expressed an interest in participating and that met the existing usage criterion were asked to sign a MoU setting out obligations arising from their participation in the trial along with those of the developers and evaluators. Once the MoU was signed, pupils in range of the study were identified in each school. Given that the school recruitment process took place for an extended period over the spring-summer terms 2018, the target cohorts of pupils in each school would during this period be Year 9 pupils who it was anticipated would enter Year 10 and commence study for GCSE mathematics from September 2018. The trial will follow these pupils through Years 10 and 11 up until the point they sit their GCSEs. Teachers

teaching Year 10 pupils at the baseline, teaching the focal cohort at Years 10 and 11 are also in range of the study, and as described above will be surveyed at four occasions.

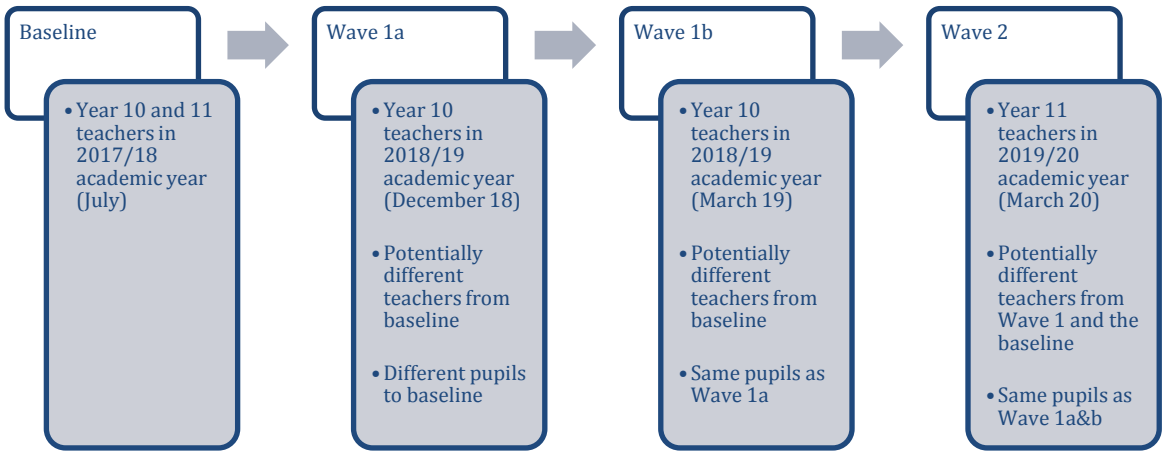

Figure 1: Sampling design for maths homework workload survey

5.1 Randomisation procedure

Once schools have signed the MoU and identified pupils within range of the trial whose parents have not withdrawn them from the study, details of all pupils, their current set for mathematics instruction and their school are sent to researchers at Manchester Metropolitan University where randomisation will be undertaken.

Schools will be randomised in batches. Within each batch schools will be arranged by region. The randomisation is stratified by region for pragmatic reasons, related to the delivery of training to intervention schools on a regional basis. Sample sizes were generally deemed to be of sufficient size such that there

was no need to stratify in order to improve sample efficiency, nor a need to use techniques such as minimisation. Within each regional stratum schools will be assigned a random number in SPSS v24 and ranked in descending order on the basis of this random number. The schools will then be divided in half with those in the lower portion of the stratum assigned to the intervention, whilst those in the upper portion to control. In strata with odd numbers of schools, the school with the largest random number will be set aside and randomised separately. This process will be repeated for each batch of schools.

6. Analysis and sample power

The analysis will be undertaken on an intention to treat basis. The purpose of the primary analysis is to obtain an effect size consistent with 'Hedges g' for the impact of Eedi on GCSE mathematics point score. Adjusted and unadjusted differences in mean GCSE scores between intervention and control group members will also be reported.

As randomisation will occur at the level of the school and GCSE mathematics point score obtained at the pupil level the data are clustered. More precisely pupils (Level 1) are nested within classes (Level 2) and classes within schools (Level 3), which means that there are three levels in the data. For this reason, effect sizes in the primary analysis will be obtained from a three level hierchical linear model estimated in STATA v15 statistical software with random effects at levels 2 and 3 in the data. In the primary analysis, two versions of the model will be estimated set out below:

$$Y_{ijk} = \beta_0 + \beta_1 t_k + \theta_k + \delta_{jk} + \varepsilon_{ijk} \text{.....[1]}$$

$$Y_{ijk} = \beta_0 + \beta_1 t_k + \beta_2 x_{ijk} + \beta_3 s_k + \theta_k + \delta_{jk} + \varepsilon_{ijk} \ldots\ldots\ldots [2]$$

The first model provides an unadjusted analysis, the second an adjusted analysis through the inclusion of the pupil level baseline test score, in this case mathematics attainment for pupils at Key Stage 2 (aged 11) in the model. Y_{ijk} represents the score at GCSE mathematics for child i in class j and school k, whilst t_k is coded one if school k is assigned to the intervention, zero otherwise. As a result, β_1 is the estimated treatment effect in the units of measurement for the dependent variable in this case GCSE score in both models. In Model 2, x_{ijk} represents the baseline test score for child i in class j and school k and s_k which captures stratification by both region and batch in which school k was randomised. In both models θ_k, δ_{jk} and ε_{ijk} represent random effects at the school, class and pupil levels with associated variances σ_k^2 , σ_j^2, σ_i^2. The effect size for impact of exposure to Eedi on GCSE score is defined as β_1 from the adjusted model (Model 2) divided by the square root of the sum of variances from the unadjusted model (Model 1) multiplied by the Hedges g adjustment factor (see Durlak, 2009). The confidence interval for this estimate will obtained on the basis of implementing bootstrap procedures (Hox, Moerbeek, & Van de Schoot, 2017). A similar approach will be taken in order to estimate effect sizes by sex and Free School Meals.

The proposed trial design will yield pre and post-intervention measures of teacher mathematics homework related workload obtained from teacher workload surveys. These data will be used to estimate the effect of Eedi on teacher workload. Estimated effects will be obtained from a two-level hierchical linear model with random effects at the teacher and school levels. Adjusted

mean differences in workload measured in hours/minutes per week will be reported with the statistical model from which estimates will be obtained containing covariates representing teacher baseline workload measure and stratification.

6.2. Sample power

Discussion of sample power focuses on the primary analysis described above. Various sampling designs are set out in Table 1 with their associated minimum detectable effect sizes (Dong & Maynard, 2013) based on a range of assumptions. The assumptions upon which the calculations in Table 1 are based draw on information obtained primarily from the Education Endowment Foundation but also other research summarising relevant empirical estimates (Bloom, 2006; Hedges & Hedberg, 2013). Standard assumptions around acceptable Types I and II statistical errors are made – namely a 5 per cent Type 1 error rate and 20 per cent Type II error rate. All statistical tests will be performed on a two-tailed basis. Crucially assumptions need to be made regarding intra-class correlation coefficients at the class and school level and the proportion of variance explained from the inclusion of KS2 points score in mathematics as a covariate in the adjusted analysis described previously (see equation [2] above).

Turning first to assumptions relating to intra class correlation coefficients. Analysis of results from previous EDUCATION ENDOWMENT FOUNDATION trials suggest intra class correlation coefficients for GCSE maths outcomes in the region of 0.15 at the school level (Allen, Jerrim, Parameshwaran, & Thomson, 2018) . On this basis, and taking a conservative approach, we adopt an estimate

of 0.20 for the crucial estimate of the intra class correlation at the school level in our sample size calculations. Further, we also allow for some intra class correlations in outcomes across classes of 0.05. Values for the proportion of variance explained through inclusion of a covariate capturing pupils score at KS2 mathematics were obtained from analyses also provided by the EDUCATION ENDOWMENT FOUNDATION (Education Endowment Foundation, 2013). Evidence suggests a correlation coefficient of around 0.7 for the association between KS2 and GCSE mathematics scores. This implies variance explained of around 0.5. We allow for some gains in precision at the level of school through the inclusion of covariates drawing on evidence provided by Bloom (2006) and Hedges & Hedberg (2013) of 0.25 variance explained.

Table 1: Minimum detectable effect sizes – whole sample estimates for primary analysis – Intention to treat

Schools	100	140	180	220
Pupils	16,800	23,520	30,240	36,960
Minimum Detectable Effect Size	0.23	0.19	0.17	0.15

Notes: Average class size assumed to be 24 pupils, with on average seven maths sets per school in Year 10 (168 Year 10 pupils in range of the trial in each school). Schools assigned 1:1 to treatment control; alpha level 0.05; two-tailed test, power 0.80. Intra class correlation coefficient at level 3 (school) assumed to be 0.20 and at level two 0.05 (class). Proportion of outcome and variances explained by covariates assumed to be .50 at level one (pupil level), 0 at level 2 (class level – we assume no class level covariates), and 0.25 at level 3 (school level). Calculations are performed using PowerUp: http://repository.upenn.edu/cgi/viewcontent.cgi?article=1265&context=gse_pubs

The choice of target sample size was driven by two considerations: 1) the fact that the intervention is relatively low cost for schools and therefore quite a modest effect size might imply a positive return on investment – this suggests a larger sample size consistent with a relatively modest minimum detectable effect

size; and 2) the cost of training teachers and liaising with schools rises as the sample size increases, given the larger absolute number of schools randomised to the intervention. In order to contain costs this suggests a lower sample size. In discussion with the study sponsors, the research team taking in account these opposing factors arrived at a target sample size of some 180 schools consistent with an effect size of 0.17. As Hattie (2008) shows, effect sizes of less than 0.20 are considered small in the context of studies in education.

7. Project team

The project evaluation team is led by Mr Andrew Boyle at AlphaPlus consultancy and Professor Stephen Morris of Manchester Metropolitan University (MMU). Stephen Morris is supported by Mr Andrew Smith (Research Associate) responsible for random allocation and sample management and Dr Zsolt Kiss, data control and statistical analysis (Visiting Research Associate) at MMU. Andrew Boyle is assisted by Dr Hayley Limmer, Dr Kathy Seymour and Clare Dowland at AlphaPlus Consultancy

8. Timetable

Date	Activity
2nd November 2017	First set up meeting, evaluation design and revisions, agreement of costs
6th December, 2017	Second set up meeting, evaluation design and cost revisions
6th-18th December	Theory of change development and agreement
2nd January – 18th May, 2018	Development of protocol, sample size, outcome measures, confirmation of data sources, randomisation approach agreed
5th February to 30th June, 2018	Recruitment of schools, MoUs signed, parental withdraw
7th June, 2018	First batch of schools randomised

late-June, 2018	Second batch of schools randomised
Early-July	Final batch of schools randomised
June-July, 2018	Training in Eedi delivered in intervention schools
July, 2018	Pre-intervention baseline teacher survey
October, 2018	Intervention commences
December 2018	Teacher survey follow-up 1
March 2019	Teacher survey follow-up 2
March 2020	Teacher survey follow-up 3
May/June 2020	Intervention ends - Focal students sit GCSEs
April 2021	Obtain NPD data extracts
Summer 2021	Analysis and reporting

9. Ethical considerations

Both AlphaPlus and Manchester Metropolitan University have ethical clearance procedures that have been invoked separately. Ethical matters in relation to this study are informed by the following considerations: 1) that the benefits or otherwise of Eedi remain unknown and have not been demonstrated – therefore preventing access to the platform would not be to remove or prevent access to a demonstrably beneficial intervention; 2) that participation of the school is on the basis of informed consent and explicit agreement, and that the consequence of allocation to either intervention or control groups are communicated to schools unambiguously prior to joining the trial; 3) that the intervention is available to schools outside of the study, therefore no school is ultimately denied access to the intervention unless they consent, and that if a school subsequently withdraws from the trial access to Eedi is restored; and 4) the trial has wider public value and therefore its design and organisation should be such as to maximise the chances of clear, unambiguous findings. The main ethical issue faced by the study was the withdrawal and prevention of access to Eedi among the control group

In the case of Manchester Metropolitan University (MMU), ethical clearance was obtained from the Faculty of Arts and Humanities Research Ethic and Governance Committee, through an expedited process, on 14 February 2018. At this stage in the research process, before the design of the trial had been fully articulated, the MMU ethics committee were told that students in the control would not be able to used Eedi for the duration of the study.

An additional ethical clearance process was initiated by AlphaPlus Consultancy as the study's principal lead organisation. This process involved clearance of drafts of the parental withdraw and information letters, data sharing and processing statement, and the school MoU. The MoU was the main means through which schools provided informed consent and signalled their agreement to participate in this trial.

Initial clearance of the MoU, parental opt-out letters and other trial documentation was received on 9th January, 2018 through AlphaPlus's ethical clearance process. At this stage, the draft MoU stated that access to Eedi for schools allocated to control would be frozen at levels of existing usage at the point the school entered the trial . Subsequently, researchers for technical and research design reasons decided that access to all existing Eedi accounts should be barred for schools in the control group. The decision was informed partly by the low levels of existing usage observed in Eedi administrative systems across the study sample but also for reasons of technological feasibility and to avoid contamination. The final version of the MoU that was used with schools made it clear that it would not be possible for control schools to set-up new nor operate

existing accounts on being allocated to the control group. It was upon this basis that informed consent was obtained from schools. It was felt by the research team that with holding all access to Eedi for schools in the control was acceptable due to informed consent received from participating schools and due to the lack of existing evidence indicating Eedi's effectiveness.

A data sharing agreement setting out the legal basis for data capture and processing was developed and agreed between AlphaPlus, Manchester Metropolitan University and Eedi/BIT. Parents are able to withdraw their children from the study at any time and can do so through informing the project team of their wish to withdraw through the school or via an online link provided in the parental withdraw letter and study information sheet.

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 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

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5–25.

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., Bos, J. M., & Lee, S.-W. (1999). Using Cluster Random Assignment to Measure Program Impacts: Statistical Implications for the Evaluation of Education Programs. *Evaluation Review*, *23*(4), 445–469. http://doi.org/10.1177/0193841x9902300405

Dong, N., & Maynard, R. A. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness*, *6*(1), 24–67. http://doi.org/10.1080/19345747.2012.673143

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology*, *34*(9), 917–928.

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

Education Endowment Foundation. (2016). *The EEF at 5*. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/public/files/Publications/5th_Anniversary_Brochure_Final.pdf

Elliott, V., Baird, J.-A., Hopfenbeck, T., Ingram, J., Thompson, I., Usher, N., … Coleman, R. (2016). *A marked improvement: A review of the evidence on written marking*. London, Education Endowment Foundation.

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, & 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.

Higton, J., Leonardi, S., Richards, N., Choudhoury, A., Sofroniou, N., & Owen, D.

(2017). *Teacher Workload Survey 2016 Research report*. London, Department for Education.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Independent Teacher Workload Review Group. (2016). *Eliminating unnecessary workload around marking Report of the Independent Teacher Workload Review Group*.

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.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice*, *30*(4), 28–37.

Konishi, C., Wong, T. K. Y., & Tao, X. (2018). Teacher support in learning: Instrumental and appraisal support in relation to math achievement. *Issues in Educational Research*, *28*(2), 202.

McMillan, J. H., Venable, J. C., & Varier, D. (2013). Studies of the Effect of Formative Assessment on Student Achievement: So Much More is Needed. *Practical Assessment, Research & Evaluation*, *18*.

Pinger, P., Rakoczy, K., Besser, M., & Klieme, E. (2018). Implementation of formative assessment–effects of quality of programme delivery on students' mathematics achievement and interest. *Assessment in Education: Principles, Policy & Practice*, *25*(2), 160–182.

Rakoczy, K., Pinger, P., Hochweber, J., Klieme, E., Schütze, B., & Besser, M. (2018). Formative assessment in mathematics: Mediated by feedback's perceived usefulness and students' self-efficacy. *Learning and Instruction*.

Raudenbush, S. W. (2008). Advancing educational policy by advancing research on instruction. *American Educational Research Journal*, *45*(1), 206–230.

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/

Shute, V. J., & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education. *Journal of Computer Assisted Learning*, *33*(1), 1–19.

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research*, *85*(4), 475–511.

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