INTERVENTION	Visible Classroom
DEVELOPER	University of Melbourne
EVALUATOR	The Behavioural Insights Team (BIT)
TRIAL REGISTRATION NUMBER	ISRCTN14774597
TRIAL STATISTICIAN	Michael Sanders
TRIAL CHIEF INVESTIGATOR	Michael Sanders
SAP AUTHOR	Michael Sanders, Hazel Northcott and Daniel Carr
SAP VERSION	2.0
SAP VERSION DATE	05.06.2018
EEF DATE OF APPROVAL	
DEVELOPER DATE OF APPROVAL	

Introduction

The intervention aims to improve student attainment in English and Maths by supporting teachers' professional practice development. The programme works by providing teachers with personalised feedback on their teaching. This feedback is designed to encourage teachers to reflect critically on their teaching and develop their classroom practice, which would be expected to have flow on effects for their pupils' learning and attainment.

The Visible Classroom (VC) intervention involves teachers' audio recording lessons and receiving detailed feedback on their teaching practices. Teachers use a smartphone or tablet app to record lessons and upload them to their own personal profile. Once uploaded, teachers receive a transcript of their lesson along with some high-level descriptive statistics of their lesson. Having uploaded a comprehensive amount of recordings (typically five hours per week), the recordings are then analysed using a teaching rubric. This contains 16 variables which are first scored individually, then converted into a composite score, to measure overall performance on the rubric. Feedback is provided to teachers based on this analysis.Teachers then work with mentors to reflect and develop their practice.

The intervention was developed by a team from the University of Melbourne and was delivered in collaboration with two partners. Ai Media provides the technological platform for the captioning of lessons, the verbatim lesson transcripts, and populates the data dashboard with teaching analytics. The Schools, Students and Teachers network (SSAT*)* are responsible for recruiting schools, checking that schools were using the technology, and supporting them to do so effectively. The University of Melbourne delivers the training package to participating teachers and mentors, conducts the in-depth coding of lesson transcripts, and generates the tailored feedback reports.

The evaluation will examine the impact of the programme on reading and mathematics attainment, measured using KS2 SAT performance in Reading and Maths. The initial intervention was designed as a two-armed randomised controlled trial involving 140 primary schools, with the assumption that we would have an average of 1.5 classes per year group and an average class size of 28 students (42 students in each school cluster). 70 schools were to be allocated to receive the intervention and 70 to a business as usual control group. Within each school, it was expected that both year five and six students would participate, to examine the impact after one year (year 6) and after 2 years (year 5) (i.e. there would be 42 students in year 5 and 42 in year 6 per school). The recruitment target was not met (86 schools were recruited) with 7156 pupils ultimately enrolled in the trial. The impact of this on power calculations is detailed in the 'Recruitment update' section.

Recruitment began in Summer 2016 and the intervention started in late 2017.

The primary research question this evaluation seeks to answer is:

- Does the VC intervention increase the educational attainment of students as measured by a combined KS2 maths and reading score for Year 5 (2 years) and for Year 6 (1 year)?

Secondary research questions will consider whether the intervention has an impact on:

- separate KS2 scores for maths and reading for pupils in Year 5 and Year 6 (analysed as separate cohorts)
- combined KS2 scores for maths and reading for students eligible for free school meals (FSM) for pupils in Year 5 and Year 6 (analysed as separate cohorts)

Study design

This is a cluster randomised controlled trial, with randomisation taking place at the school level. As the intervention entails a change in initial practice for teachers (recording their lessons), receiving feedback and then discussing with a mentor from their own school, it was felt that the risks of spillover in a class level randomisation were substantial. The trial aimed to recruit 140 primary schools (ultimately 86 were recruited), with schools randomly allocated to either the treatment arm or the control group. Schools in the control group were expected to continue with 'business as usual'. After the study completes in August 2018, schools in the control group will be offered the choice of either the intervention (at no cost) or a £1,000 payment from the EEF.

The trial is being conducted across England, with no geographic restrictions. The eligibility criteria for schools to participate were:

- Upfront transfer of data to the evaluators including eligible student UPNs;
- A completed Memorandum of Understanding;
- Agreement that teachers in both the VC intervention and control arm complete a survey at the end of the trial period;
- Inclusion of both Y5 & Y6 in the intervention, with a minimum of two teachers providing baseline data and attending the training session;
- Teachers must have access to a tablet or smartphone and sufficient internet connection to upload recordings (internet connection to be verified at baseline data gathering);
- The schools must not be using Visible Learning plus, a similar intervention designed by the University of Melbourne Project Team.

In mid-August 2016 a decision was made during school recruitment to begin capping the number of teachers per school who could be involved in the trial (see Table 1 for details of numbers). This decision was made to avoid a cost overrun in the delivery team's budget. As capped schools may select which teachers can be involved in this trial on the basis of merit or some other element correlated with student attainment, we made the decision to include whether the school was offered capped recruitment as a covariate in our stratification.

There was no eligibility criteria for students, other than that parents did not request their child not be included in the evaluation analysis.

Randomisation

Once schools had been recruited and had signed a Memoranda of Understanding (MoU)

student data was collected. Randomisation occurred following baseline data collection. To ensure balance between the treatment and control groups in important covariates, randomisation was stratified on the basis of school level information students' FSM status, previous test scores, and whether the school was recruited into the trial on a capped or uncapped basis.. The randomisation followed a two-stage process:

1. Schools were stratified on the basis of the proportion of FSM students (split by whether this was above or below the sample median proportion), 2010-11 KS1 Average Point Score (split in a similar way), and whether schools were offered entry into the trial on a capped or uncapped basis.
2. A random number was generated within each block to ensure that the proportion of FSM students, KS1 results, and number of and capped and uncapped schools were balanced across trial arms. We used data from DfE's Performance Tables to determine the blocking characteristics, and SSAT noted which schools were offered capped entry to the trial.

A total of 86 schools were recruited, with 44 randomised to treatment and 42 to control. In Table 1 we set out the breakdown across stratification variables. In a small number of cases, data required for stratification are missing (3 schools are missing FSM data, and 2 schools are missing KS1 data).

Table 1. School stratification

Stratification Variable	Treatment	Control
Proportion FSM: above median	21	20
Proportion FSM: equal to or below median	21	21
KS1 point score: or above median	20	20
KS1 point score: equal to or below median	22	22
Capped entry	5	6
Uncapped entry	39	36

Calculation of sample size

Initial Power Calculations

These power calculations were based on the analysis of Y5 & Y6 separately. In addition we assumed that:

- Sample size= 2,940 students per arm. This was estimated on the original assumption that 140 schools participate in the trial, with an average of 1.5 classes per year group and an average class size of 28 students (42 students in each school cluster).
- Randomisation is at the school-level.
- The Intracluster correlation coefficient is 0.15.[1]
- Any attrition will be minimal. As tracking of students who change schools during the trial will be possible, through Unique Pupil Numbers (UPNs), we expect attrition to be minimal, however below we do set out a process in the instance that significant data is missing.
- Participants' KS1 results will be incorporated to increase power along with the variables used for blocking during randomisation. We assume a correlation coefficient of 0.7 between KS1 and KS2 results under the null hypothesis.[2]
- Hypotheses:
    - Null hypothesis: There will be no difference in combined maths and English KS2 scores between students whose teachers used Visible Classroom and those whose teachers did not.
    - Alternative hypothesis: There will be a difference in combined maths and English KS2 scores between students whose teachers used Visible Classroom and those whose teachers did not (i.e. a two-sided alternative hypothesis).
- Power: 80% ; Significance level: 5%. These figures are standard in social and education policy trials.
- We have 2 trial arms (a treatment and control).

Given these assumptions, we expected a minimum detectable effect size (MDES) of 0.139 (Cohen's D).

FSM power calculations

Based on the same assumptions as above, we calculated the MDES for FSM students (assuming 16.7% of students receive FSM as per the average across Primary Schools on the EduBase dataset) to be 0.175.

Recruitment update

The original recruitment target was ultimately not met. 86 schools were recruited, and many

[1] This is based on estimates from Hedges & Hedberg (2007), denoting an ICC of 0.13-0.21 for schools. Hedges, L. & Hedberg, E. (2007). Intraclass correlation values for planning group-randomized trials in education. Educational Evaluation and Policy Analysis, 29 (1), 60-87.
[2] https://educationendowmentfoundation.org.uk/public/files/Evaluation/EEF_Evaluation_Pre-testing_in_EEF_evaluations.pdf

of them quite close to the deadline for randomisation. Due to time constraints, two dozen schools were randomised into the trial prior to confirmation that opt-out forms had been distributed to pupils, under the condition that they provide a full list of students and then subsequently confirm which were to be removed (if any parents requested that their children be removed from the study analysis). During the months that followed 23 of the 24 schools were able to confirm that they had distributed opt-out forms to parents, and returned names to be removed from the study in instances where requests from parents had been received.

The school that did not confirm this during the 2016-17 academic year, due to change in staff and a loss of the forms collected, provided opt-out forms again to parents at the start of the 2017-18 term, and submitted names to remove where opt-out requests were made. Given that for this school, the Year 6 cohort had by this point already moved on to secondary, there are only 85 schools available for analysis at the 2016-17 Year 6 cohort level. The remaining school was deemed to have partially attrited, and will not be included in the 2016-17 Year 6 cohort analysis.

In terms of the number of pupils ultimately enrolled in the trial, the total number is 7156, somewhat lower than the assumptions made in the original power calculation.

Factoring in the final number of schools and students enrolled into the trial, and keeping all other assumptions the same, the MDES expected are as follows:

Table 2.

Year level	**No. of schools**	**No. of pupils**	**MDES** (Cohen's D, assuming clusters equal size)
Year 6	85	3547	0.179
Year 5	86	3609	0.177

Outcome measures

Primary outcome

The primary outcome measure will be performance in combined scores from KS2 reading and maths tests for years 5 and 6, with years 5 and 6 being analysed as separate cohorts. Scores will be combined by adding the raw scores together, which is the most straightforward means of combination and produces identical standardised effect sizes to alternatives. These will be measured using NPD extracts as follows:

- **KS2 Maths**: KS2_MATMRK (the sum of marks across KS2 mathematics papers) will be our endline measure, with KS1 maths performance as measured by KS2_KS1MATPS used as a baseline.
- **KS2 Reading**: KS2_READMRK (mark in the KS2 reading paper) will be our endline measure, with KS1 maths performance as measured by KS2_KS1READPS used as

a baseline.

Secondary outcomes

There are no secondary outcomes in this trial, only secondary analysis on the FSM subgroup.

Analysis

Primary intention-to-treat (ITT) analysis

Our primary analysis will focus on KS2 maths and reading results separately by year level, and will be performed using Stata (version 14). Outcome variables will be regressed using a least squares linear model with treatment arm indicators, strata indicators (FSM students (split by whether this was above or below the sample median proportion), 2010-11 KS1 Average Point Score (split in a similar way), and whether schools were offered entry into the trial on a capped or uncapped basis), and student-level KS1 baseline attainment as covariates. To account for the experimental design, standard errors will be clustered at the school level to allow for correlation of pupil outcomes within schools.

The estimated impacts will be intention to treat (ITT) effects and will be reported with 95% confidence intervals. Intra-cluster correlations will also be reported.

$$Y_{ijk} = \alpha + \beta_1 T_{jk} + \beta_2 Y_{i(k-1)} + \beta_3 X_{ik} + \varepsilon_{ijk}$$

where i are individuals, Y_{ijk} is our endline KS2 score (with the value of this at $k-1$ being the KS1 score for the same student), T_{jk} is our school-level treatment indicator (so T_{jk} is one if student *i* is in a school *j* which has been treated and zero otherwise), X_{ik} being a vector of stratification variables, and ε_{ijk} being an error term. Errors will be clustered at school-level. Our primary intention to treat outcome will be recovered from the estimate of β_1 when this model is estimated on the full sample at randomisation. This model will not be altered depending on the significance of any variables included (i.e. all variables will be retained in the model regardless of whether they are statistically significant) including the vector of blocking variables (X_{ik}).

Imbalance at baseline for analysed groups

We will check for balance of the analysed sample on the following characteristics:

- baseline KS1 reading and maths point scores,
- proportion of students who are female,
- proportion of students ever eligible for Free School Meals (using EVERFSM_6_P_[term][year]),
- proportion of students for whom English is an Additional Language, and
- Age of student in months

We will do this by calculating absolute standardised differences (Imbens & Rubin, 2015) between the treatment and control groups and these will be presented in the report.

Missing data

Due to the nature of the outcome variable (KS2 results), we do not foresee missing data to be problematic. Any missing data will be summarised based on the reasons for missingness, broken down by trial arm. At this stage, the most likely causes of missing data are the withdrawal by either participants or the school (as data controller) of consent, or a change in the status of the student such that their data is not available (i.e. under some circumstances if the young person is taken into care). If we observe that more than 5% of our outcome data are missing, then we will fit a regression where the outcome is the missingness of Y, and where the explanatory variables are student level prior attainment, gender, and free school meals status, and characteristics of the school including its Ofsted rating. We do not believe that a fully specified model with school level fixed effects is appropriate here as given the scale of the trial this is likely to produce spuriously significant predictors of missingness. If this indicates the missing data are systematic, we present the procedure for sensitivity analysis based on the missing data. Missing data presents a problem for analysis, whether a pupil is missing a value for an outcome variable (endline score) or for covariates (e.g. baseline score). If outcome data is 'missing at random' given a set of covariates then the analysis has reduced power to detect an effect; if data is 'missing not at random' (for example, differential dropout in the intervention and control groups for unobserved reasons) then omitting these pupils (as in the primary 'completers' analysis) could bias the results. Conducting sensitivity analysis through imputing missing data could improve the robustness of the analysis and examine how sensitive the results are to alternative assumptions.

Every school that will be randomised will complete an endline in the form of KS2. Though some individual pupils within those schools may have left the school, we will still be able to collect their data via the NPD. However, KS2 attainment data will be missing for pupils absent during their exams. Where >5% of the total sample is missing KS2 attainment data, and where our analysis described above suggests that data are missing at random, we will perform sensitivity analysis and impute the missing scores for those pupils using multiple imputation, including KS1 attainment, census (FSM status) and school-level data (school-level performance in the baseline data) in the model for imputation. As the correlation between KS1 and KS2 attainment is high, the imputed values will be estimated with a high level of precision. Where the results from the multiply imputed regression substantially disagree with the results of the primary analysis, a decision must be taken as to which is the more robust. Where the difference is a matter of statistical significance driven by precision of the estimate and not by changes in the point estimate, then we will revert to the primary analysis as specified above. Where there is substantial difference in the point estimates of the effect (either with or without changes in significance) between the primary and imputed analysis, we will conduct further sensitivity analysis by sequentially dropping each of the variables included in the imputation in turn. If the differences persist under this sensitivity analysis, then the multiply imputed model will be taken as the main measure of the effect.

Additional sensitivity analysis will be conducted for pupils with KS2 attainment data, but no KS1 attainment data. In these instances KS1 performance will be imputed using the schools mean KS1 score.

Pupils who are missing both KS1 or KS2 attainment data will be discarded from the analysis.

Non-compliance with intervention

The following criteria have been defined in the trial protocol as variables that can be used to assess dosage of the intervention. These have been devised in collaboration with the University of Melbourne and can be defined as follows:

To be considered to have been treated (i.e. a complier), a teacher must have received at least 2 reports from Visible Classroom, which means they must have submitted at least 10 recordings over the academic year.

We will use Complier Average Causal Effect (CACE)[3] analysis to estimate intervention effects on treated children. We will estimate the CACE using two stage least squares (2SLS) regression by estimating a (first stage) model of compliance, using the definition of "on treatment" described above. The predicted values from the first stage are then used in the estimation of a (reduced form) model of our outcome measure □□□. In other respects, the specification remains the same as the primary outcome ITT model. We will conduct this analysis using the `ivregress` functionality of Stata to make necessary adjustments to standard errors (which will also be clustered at school level) due to the instrumental variables approach. We note the need for caution in generalising the results of this analysis, given the likely endogeneity of attendance and motivation.

Secondary outcome analyses

We will repeat the primary analysis using individual scores for the KS2 attainment measures, with analysis for years 5 and 6 conducted separately.

Subgroup analyses

We will also conduct an analysis for the FSM subgroup of pupils, using the same model as our primary analysis (i.e. combined maths and reading separately for Year 5 and Year 6 - specifically those who are registered for free school meals (FSM) in the National Pupil Database (using the variable EVERFSM_6_P, following EEF guidance).

This subgroups was identified in the trial protocol. FSM pupils are clearly a key subgroup to be analysed in all EEF trials. For FSM pupils, analysis will be run separately for this subgroup, in line with EEF analysis guidance.

Additional analyses

Ideally, we would test whether there is a smaller (or larger) effect of the intervention for teachers who score on the bottom third of the VC rubric, as the university of Melbourne hypothesise that a different effect size is likely. However, because performance rated by the rubric is not available for the control group, we do not have a counterfactual for either the higher or lower performing teachers as rated by the rubric. Given that high or low performance on the rubric is likely to be correlated with other, unobserved characteristics such as teacher quality, motivation, and engagement with the trial, it is not possible to

[3] Gerber AS, Green DP. (2012) Field Experiments: Design, analysis and interpretation. WW Norton and Company, New York.

attribute any differences between high and low performing teachers to the intervention itself and any causal analysis is likely to be misleading.
Instead we will conduct a descriptive dose response analysis, in which our treatment variable is pseudo-continuous based on the VC rubric performance of treated teachers.

In addition, we will conduct analysis using the Survey of Teacher Practice to understand how self-reported changes in teaching practice moderate any effect on attainment. As per the Trial Protocol, the Survey of Teacher Practice will be collected for all teachers.

Effect size calculation

Effect sizes will be calculated in line with the EEF's analysis policy for cluster randomised trials i.e. using total variance (rather than within cluster variance) to maximise comparability with other trials. This will require estimates of:

- the unstandardised treatment effect ($\hat{\beta}_1$) from the primary ITT analysis regression model reported above;
- the total standard deviation for the analysis sample ($\hat{\sigma}_T$), which can easily be recovered from the estimation model. (Note that $\hat{\sigma}_T$ is a combination of variance within schools –$\hat{\sigma}_W$–and between schools – $\hat{\sigma}_B$– although there is no need to decompose given that estimation is not carried out using a hierarchical model.)

$$ES = \frac{(\underline{\hat{\beta}_1} - \underline{\hat{\beta}_2})}{\hat{\sigma}_T} = \frac{(\underline{\hat{\beta}_1} - \underline{\hat{\beta}_2})}{\sqrt{\hat{\sigma}_W^2 + \hat{\sigma}_B^2}}$$

where our estimate of $\underline{\beta_1} - \underline{\beta_2}$ is recovered from $\hat{\beta}_1$in the primary ITT analysis model.

Ninety-five percent confidence intervals (95% CIs) will be estimated by inputting the upper and lower confidence limits of $\hat{\beta}_1$ from the regression model into the effect size formula.

Report tables

We do not anticipate including tables in addition to those expected as per EEF reporting guidance.