INTERVENTION	University of Bristol Teacher Observation
DEVELOPER	CMPO, University of Bristol
EVALUATOR	NFER
TRIAL REGISTRATION NUMBER	ISRCTN89620259
TRIAL STATISTICIAN	Jack Worth
TRIAL CHIEF INVESTIGATOR	Ben Styles
SAP AUTHOR	Ben Styles
SAP VERSION	4
SAP VERSION DATE	2/11/16
EEF DATE OF APPROVAL	16/9/16
DEVELOPER DATE OF APPROVAL	

Contents

Introduction

The teacher observation intervention is being delivered by CMPO (Centre for Market and Public Organisation) by principal investigator Professor Simon Burgess, using funding from the Education Endowment Foundation. The programme has two main aims: to improve teacher effectiveness and to improve learners' educational outcomes. It seeks to achieve these aims through teachers observing each other and being observed themselves. Observations are planned to occur a large number of times over the course of a year. They will take place in maths and English departments across all intervention schools and using a tablet with RANDA software to record the observations. The impact of the intervention on learners' ability will be measured by their GCSE mathematics and English results and their attainment at the end of year 10 in bespoke tests developed by NFER.

Study design

A sample of secondary schools were approached who are nationally representative (excluding Somerset, Merseyside and Lancashire) from schools with the highest percentages of pupils on free school meals (FSM). The 92 recruited schools were then randomly assigned to one of two groups (41 intervention schools and 41 control schools; 10 withdrew without knowledge of group allocation):

- Teacher peer observation (referred to subsequently as 'intervention')
- 'Business-as-usual' control (referred to subsequently as 'control')

Some teachers are observers, some observees and some observe and are observed (a third of teachers in each group). The number of observations received should vary - either 6 a year (low observation category) or 12 a year (high observation category). English and maths departments in each intervention school will be randomly assigned to each dosage so every school has one low and one high observation category. Within these departments, teachers will be randomly assigned to observer/observee/both. The pilot revealed that it will not be possible to specify the number of observations carried out by those in the observer or both categories due to the common practice of schools timetabling all English/maths lessons at the same time. Instead, the intended minimum number of observations carried out will be 3 in the low dosage departments and 4 in high dosage departments.

Protocol changes

The randomisation procedure was changed from minimisation to stratified randomisation. The strata used, however, were the same so this does not impact on analysis.

Randomisation

82 schools have been randomly allocated to intervention (41 schools) and control (41 schools) groups using stratified randomisation. The strata used were school performance, eligibility for free school meals (FSM) and ethnic background. These were all calculated as binary variables with high/low options. School performance was calculated by taking 2013 scores for school maths VA (maths KS2 to maths GCSE accounting for student gender, major ethnic group and FSM), and school English VA (English KS2 to English (language) GCSE accounting for student gender, major ethnic group and FSM). As these are generally highly correlated they were combined to make a single variable with two groups- high and low performance. Free school meals was calculated by percentage of students eligible for free school meals in the school, split by the median. Ethnicity was percentage of white students in the school split by the median.

Originally 92 schools were randomly allocated to groups but ten schools did not supply the student data by the deadline given, which was an inclusion criteria outlined within the protocol and therefore were not informed of their group allocation and were withdrawn from the trial. Within the

intervention group schools the English and maths departments were randomly allocated to one of high dosage and the other low dosage. The teachers were randomised to one of three groups; observer, observee or both. If a teacher leaves the trial and is replaced one-for-one, the replacement teacher continues in the role the original teacher was randomised to. If a school takes on a new teacher in addition to existing Year 10 and Year 11 staff (or the number taken on does not equal the number who leave), this teacher is randomised to one of the three observation groups.

Calculation of sample size

The protocol power calculations recommended the following recruitment thresholds:

Total number of schools recruited	**Action**
>= 50	Proceed with school randomisation only (high dosage only and all teachers both observers and observees)
>=70	Proceed with school and teacher randomisation (high dosage only)
>=100	Proceed with all three experiments

After discussions with the developer and the funder, it was agreed to continue with all three experiments despite not achieving the 100 recruitment threshold. This was done with the knowledge that the departmental randomisation will dilute the intervention, thus reducing the possible ES and therefore power.

Subsequent to the protocol power calculations, it is noted that the assumed intra-cluster correlation (ICC) for the teacher-level experiment (0.075) is likely to be an under-estimate due to the widespread practice of setting. Setting is highly prevalent in secondary schools, particularly in mathematics, so it may be that the larger n in the teacher experiment is undermined by an excessively large ICC. This will be mitigated in part by the baseline measure used as a covariate in each analysis model.

Outcome measures

Primary outcome

The primary outcome for the school-level experiment will be mathematics and English GCSE outcomes combined and equally weighted for Year 11 students who have been involved in the trial for two years i.e. those that started Year 10 in 2014.[1] Specifically, and in terms of the September 2015 edition of NPD Data Tables[2], KS4_EBPTSMAT_PTQ (Point score in maths EBacc pillar) will be added to KS4_EBPTSENG_PTQ (Point score in English EBacc pillar) to create the primary outcome. It is anticipated that the number of students with only one of these outcomes is likely to be very small so these will be excluded from the analysis. In the event that this number is greater than 5% of cases, see Analysis section below.

[1] Note that the protocol refers to two separate primary outcomes on the basis of this being necessary for the dosage analysis. For the main school-level experiment, we require a single primary outcome to avoid the problem of multiple inference.

[2] https://www.gov.uk/government/publications/national-student-database-user-guide-and-supporting-information

Secondary outcomes

The secondary outcomes will be the total scores of the year 10 bespoke tests in English and maths. Year 10 test results will be analysed at the end of both the first and second years of the trial. Other secondary outcomes will be the individual point scores in each of maths and English at Year 11 at the end of both the first and second years of the trial.

Analysis

School-level experiment

The primary outcome analysis will be 'intention to treat' (ITT). An interim analysis of the Year 10 data (a secondary outcome) from the first year of the school-level experiment indicated that a multi-level model with two levels (school and student) was preferred to one with three (school, teacher and student). This was because approximately 19% of student-level degrees of freedom were lost from the three-level model due to the imperfect coverage of the teacher-student linked lists and the additional random effects due to the extra level. KS2 test result (sum of KS2_MATTOTMRK [Total marks achieved in Maths test (sum of Paper A, Paper B and mental arithmetic tests)] and KS2_ENGTOTMRK [Total marks achieved in English test (sum of reading and writing tests)]) will be used as a covariate in the primary outcome model[3]. As per the updated EEF analysis guidelines (December 2015) no further covariates will be included aside from the three school-level variables that were used to stratify the randomisation. All four covariates will be entered into the model regardless of whether they are significant. The R package nlme will be used to run the multi-level model.

The primary analysis will be on 'complete' NPD obtained for all randomised schools that were alerted to their group allocation i.e. n=82. Of the original 92 randomised, 10 schools dropped out of the trial before allocation was known; their dropout can be considered unbiased so these schools will not be included in the ITT analysis.

Missing data is unlikely to be a problem for the primary outcome analysis as it is obtained from NPD. Missing data generally presents a problem for analysis, whether a pupil is missing a value for an outcome variable (post-test score) or for covariates (e.g. pre-test 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. Imputing missing data could improve the robustness of the analysis and examine how sensitive the results are to alternative assumptions. It can also signal missing not at random if the imputed result is much different from the completers analysis. Likelihood-based methods (e.g. nlme function in R) are usually consistent with the results from multiple imputation if the missingness mechanism is missing at random.

A discussion of the results in the context of missing follow-up data will be presented. If follow-up data is missing at random given covariates, and these covariates are included in the model, the results will be unbiased. If greater than 5% of cases have missing baseline data as compared to the definitive student list, multilevel multiple imputation will be used (see www.missingdata.org). It may be that the results of the multiple imputation do not differ appreciatively from the completers analysis. If this is the case and we are reasonably confident that covariates explain any missingness then this will complete the primary analysis. Otherwise, some sensitivity analysis (e.g. using extreme values) may be necessary.

The primary analysis will be followed by an 'on-treatment' analysis where RANDA data from the tablets will be used to determine the extent of each teacher's involvement and will replace the

[3] In the protocol, KS3 teacher assessments were going to be used as a covariate as they correlate more highly with GCSE grades, however, these are no longer available on NPD so cannot be used.

intervention group variable in the model. This analysis will enable us to estimate a 'pure intervention effect' (net of any fidelity issues) that is not necessarily causal in nature.

Secondary outcome analyses will mirror that of the primary outcome but only the corresponding subject's KS2 score[4] will be included as a covariate, alongside the stratification variables.

Department-level experiment

This experiment is being carried out within the intervention group of the main school-level experiment. To avoid the proliferation of secondary analyses and because this is the lowest powered of the three experiments, only Year 11 data from the second year of the trial will be used. Half the maths departments were randomised to high dosage and the other half to low dosage. We will use a multilevel model with two levels (school and student) to model point score in KS4 maths using point score in KS2 maths as a covariate. No further covariates are required as there were none used in the randomisation of departments. The English department experiment will be modelled in the same way.

The primary analysis will be on 'complete' NPD obtained for all randomised schools that were alerted to their departmental allocation i.e. n=41. Five schools that were eligible for the departmental experiment dropped out of the trial before allocation was known; their dropout can be considered unbiased so these schools will not be included in the ITT analysis.

Teacher-level experiment

We will need to establish if a learner has one teacher for each subject during the course of the year of study[5]. If in fact each learner has more than one teacher then there is the possibility that they will be receiving more than one strain of the intervention (for example having one teacher who is an observer and one is who is an observee) which could change the impact of the intervention. In addition, to be able to accurately test the effect of both dosage and being an observer or observee we need to assume that activities in English and mathematics do not influence each other in terms of attainment. Teachers' perceptions on this will be explored during the process evaluation.

Assuming we are able to allocate each student to a single teacher (i.e. the teacher that has had the most contact over the course of two years), the maths teacher experiment will be analysed using a three-level (school, teacher and student) multilevel model of KS4 points score in maths with KS2 points score in maths as a covariate. The randomisation for this experiment was stratified by school and department. As the analysis will be by subject, the department stratification is covered by including school as a level in the model. The English teacher experiment will be modelled in the same way. If it is common for students to be allocated to more than one teacher, a cross-classified multilevel model may be required.

This experiment has a factorial design as the 'observer' and 'observee' categories overlap for the 'both' category. However, it does not contain all combinations of factors as no teachers were randomised to a 'do nothing' category. We will therefore model the data using two factors but without their interaction (see Table 1).

Table 1. Values of factors in the teacher experiment

Type of teacher	**Factor 1**	**Factor 2**
observer	1	0
observee	0	1
both	1	1

[4] The 2010 KS2 boycott may affect the 2015 Year 11 analysis. If so, a measure that incorporates teacher assessment may be needed.
[5] And over two years in the case of the year 10 cohort that starts the trial in October 2014.

The majority of students did only have one teacher per subject at Year 10. When there were two teachers, if it is not clear which teacher had the most contact, sensitivity analysis will be performed using, for example, the second teacher ID in place of the first, where it exists.

The primary analysis will be on NPD data obtained for all randomised teachers. Teachers in the 41 intervention schools were randomised by NFER but, before the results were communicated to CMPO, a further three schools dropped out of the study. As this occurred without knowledge of group allocation, this can be considered unbiased attrition. This experiment will hence be analysed with data from the remaining 38 schools whose teachers were randomised.

Subgroup analyses

Sub-group analysis on the primary outcome will be carried out on the following groups only as per the protocol: gender and whether or not a pupil has ever received free school meals (everFSM). This will be done using a model identical to the primary outcome model but including gender, everFSM, gender*intervention and everFSM*intervention as covariates. A separate primary outcome model (with no extra covariates) will also be run on everFSM students alone as per all EEF trials.

Effect size calculation

All effect sizes will be calculated using total variance from a multilevel model, without covariates, as the denominator i.e. equivalent to Hedges' g. The numerator will be the raw coefficient for the intervention group from the multilevel model. They will be reported with a 95% confidence interval that takes into account the clustered nature of the data. The upper and lower bounds of the confidence interval will be calculated as the effect size plus/minus the product of the critical value of the normal distribution (≈ 1.96) and the standard error of the effect size estimated from the multilevel model.

We have deliberately kept the analysis of each experiment true to its randomisation. This has the advantage of limiting the number of comparisons that could be used to justify that the programme has 'worked'. The first experiment should be the judge of this and subsequent experiments unpick what is going on within the 'black box'. Note that if the control schools from the first experiment were included in the analysis of subsequent experiments, any one of a large number of analyses might be construed as demonstrating success. Such conclusions would be undermined through the family-wise error rate.

Further analyses for report

- Sample representation analysis
- School characteristics – of 82 schools post randomisation
- Student characteristics –fsm, gender and KS2 scores
- Histograms of year 10 test performance at year 1 and year 2; Cronbach's alpha for each test to indicate reliability
- MDES calculation – on the basis of actual parameters seen
- Baseline effect size – multilevel model of baseline score (KS2) against intervention group indicator for those students in the final model to determine whether attrition has led to a significant imbalance at pre-test
- Student characteristics of analysed groups – ANOVA by intervention group of school-level background factors percentage female, percentage everfsm; to check for possible bias introduced due to attrition.