# Teacher Observation Intervention

NFER

Ben Styles

Education
Endowment
Foundation

## Amendments

Following development work in the pilot several changes have been made to the original protocol as detailed in full below:

- Page 2- more detail has been added about the sampling exclusions including the geographical areas and the detail about FSM schools- these exact specifications had not been made when the original protocol was developed. The FSM criteria are referred to again at the bottom of page 2.
- Page 2- details of low dose and high dose specifications for observers and obeservees have been added- these had not been exactly specificed before the pilot and were informed by the findings.
- Page 2- inclusion criteria has been changed to state that UPNs, class lists and anonymous teacher ids will be collected from schools rather than teacher names as CMPO felt that this would encourage schools and teachers to participate if they would remain anonymous.
- Page 2- consent has been updated from opt- out parental consent letters to headteacher consent through signed memorandums of understanding after this was approved by NFER's code of practice committe as the most appropriate method of consent.
- Page 2- the method of randomisation was updated from minimisation to stratification following discussions between NFER and CMPO about the most appropriate methodology to ensure balance across key strata identified by CMPO based on previous research evidence.
- Page 3- choice of year 10 tests was confirmed as adapted versions of Key Stage 3 tests- this was suggested in the original protocol but had not been confirmed.
- Page 5- a table of recruitment thresholds was added with new power calculations in the event of a reduced number of schools participating in the trial.
- Page 5- more clarity was added about the way that Year 10 pupils will be randomnly allocated to either sit the maths or English exam
- Page 6- figures have been removed from the sentence referring to numbers of observations in the high and lose dose as this is now been established as a range rather than a single number.
- Page 6- Juliet Sizmur is identified as the personnel responsible at NFER for the Year 10 testing as this was not specified in the earlier protocol.

| Evaluation Summary | |
|---|---|
| **Age range** | Secondary |
| **Number of pupils** | 18,000 |
| **Number of schools** | 120 |
| **Design** | Randomised controlled trial with randomisation at the school level |
| **Primary Outcome** | English and Maths |

## Protocol for evaluation of the teacher observation intervention

## Intervention

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 by teachers observing and being observed a large number of times by peers over the course of a year in maths and English departments in 120 schools and using a tablet with RANDA software to record the observations. A similar structured lesson observation approach has been tested in the US[1]. This showed that students performed better in the year that their teachers were undergoing the structured observation and even better in subsequent years. The effect on student attainment was relatively small, at 0.11 standard deviations for mathematics attainment but this needs to be tested using a RCT. It is anticipated that the impact of the study will continue to develop over time, perhaps as schools adopt a cultural shift to observation and therefore it is recommended that all schools are advised at the start of the intervention that although the trial is only two years we may want to review NPD data after this time and the intervention group may want to continue with the intervention method after this time (but this is not included as part of this trial or in these costs).

## Research Plan

### *Pilot*

The research design will be piloted in 3 schools by CMPO. The purpose of the pilot will be to:

1. Test how the RANDA software and rubric work and identify/solve any technical issues for delivery- CMPO
2. Gain an understanding of the technical requirements for the software- e.g. broadband- CMPO
3. Carry out a small scale process evaluation involving case study visits to the schools involving interviews with senior leaders and those taking part in the pilot to check feasibility of the intervention including gathering feedback on whether the varying dosages of observations are appropriate and how the split between observers and observees has worked- NFER
4. Test the design and randomisation- whether the department split works; whether the learner/ teacher level information is correct in respect to who is teaching which learners; and if the split within a department for the observation type works- NFER
5. Trial/pilot the year 10 tests, with a view to exploring and refining the reliability and validity of the assessments to be used in the main trial- NFER.

At the end of the pilot we will have a meeting with partner organisations and produce a short summary report of any challenges with the pilot and recommendations of any changes for the main trial.

[1] Taylor, Eric; Tyler, John (2012) Can Teacher Evaluation Improve Teaching? Education Next, 12, 78-84.

## *Research Questions and overall design*

The primary research question is: what is the impact of teacher peer observation on learners' GCSE mathematics and English achievement?

The secondary research question is: what is the impact of teacher peer observation on learners' Year 10 mathematics and English achievement?

A sample of secondary schools will be approached who are nationally representative (excluding Somerset, Meryside and Lancashire) from schools with the highest percentages of pupils on free school meals (FSM). The 120 schools recruited will then be randomly assigned to one of two groups (55 intervention schools and 65 control schools):

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

Some teachers will be observers, some observees and some will observe and be observed (a third of teachers in each group). The number of observations received will 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 minimum number of observations carried out will be 3 in the low dosage departments and 4 in high dosage departments.

The main trial will start in September 2014 and will run until July 2016 and will use GCSE and year 10 results from July 2015 and July 2016.

The trial will be designed, conducted and reported to CONSORT standards (http://www.consort-statement.org/consort-statement/) and registered on http://www.controlled-trials.com/.

## *Inclusion Criteria*

To be included in the trial, schools will need to provide a list of UPNs and class lists linking anonymous teacher IDs to UPNs to CMPO to avoid control schools resisting the data requirements after allocation. The code of practice committee at NFER decided that headteacher consent is appropriate for this study and this will be collected by CMPO through signed memorandums of understanding. We will only be looking at Year 10 and 11 learners who are from schools with the highest percentage of students receiving FSM. At this stage we would also recommend that there are several recruitment events to explain the evaluation to schools, so they understand the importance of the randomisation methodology before signing up to the trial (NB. there will not be very much information on the intervention here to ensure that this does not affect the realistic nature of the trial or to risk contamination in the schools that are allocated to the control group when randomised).

## *Randomisation methods*

Randomisation on all levels will be carried out by a statistician at NFER. Schools will be allocated to groups (55 intervention and 65 control) using stratified randomisation to guarantee balance across a range of school-level measures. Departments within intervention schools will be randomly allocated to low or high observation categories in the ratio 1:1. Teachers in intervention schools will be randomly allocated to be observers/observees or both in the ratio 1:1:1.

It is important to note here that we will need to establish if a learner has one teacher for each subject during the course of the year of study[2]. 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

[2] And over two years in the case of the year 10 cohort that starts the trial in October 2014.

one teacher who is an observer and one is who is an observee) which could change the impact of the intervention. This will be investigated during the pilot study and will also need to be explored in the process evaluation. 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.

### *Outcome Measures*

GCSE outcomes for English and mathematics will be used to measure year 11 ability.

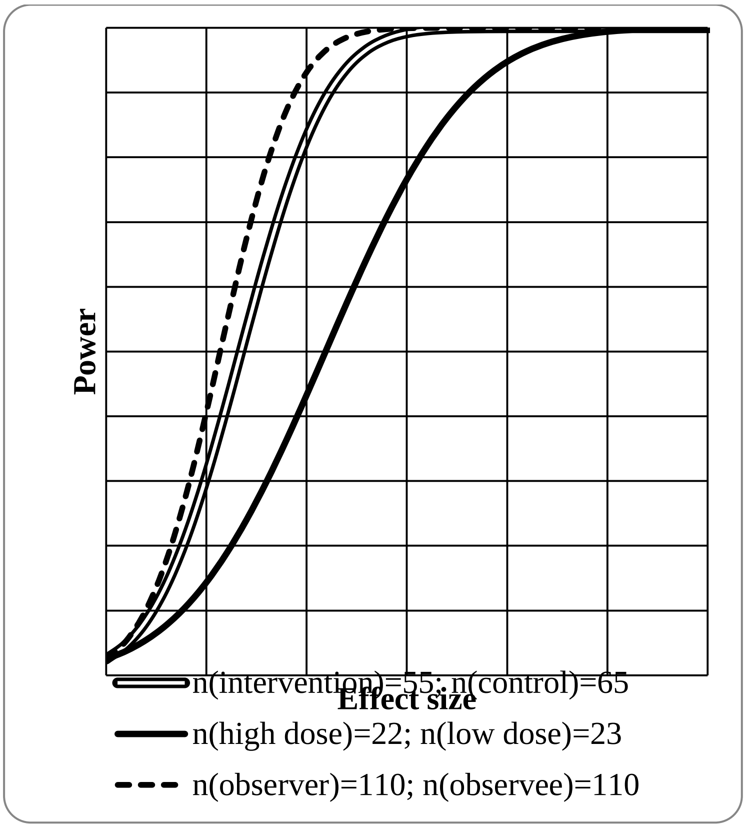
A number of the commercial tests available cover very basic skills and are not precisely tailored enough to provide a valid assessment of the types of changes in student performance that might be expected to result from this intervention. Therefore, to measure Year 10 ability, we will construct bespoke tests by selecting items from Key Stage 3 past papers. These, we know, have been robustly developed and trialled and are appropriate to the age range and the curriculum. Although the existing KS3 mathematics tests, for example, last 2 hours, we would propose selecting a range of items that address higher order skills in order to construct suitable tests for the main trial for the test to last around an hour. Standardisation data is not essential where both intervention and control groups are being tested on the same instruments.

KS3 tests are in the public domain and freely available for educational purposes (http://www.sats-papers.co.uk/sats-papers-ks3.php), but there would, of course, be printing and marking costs, as usual. Nonetheless, we believe this could be done within the budget previously discussed (£5 per test)[3] and would provide a more valid measure that would increase the likelihood of detecting a change in student performance, compared with a more 'blunt' commercially available instrument. We would include an expert review and some informal trialling of the selected items before gathering data from pilot schools. This would be followed by item analyses and reliability analyses to ensure that our item selection is pitched at an appropriate level and is addressing coherent underlying constructs.

Year 10 pupils will be randomly allocated to taking either a mathematics or English test which will be administered by NFER test administrators.

[3] Please note – this does not include the cost of test administration, which will be significant due to the extent of testing.

### *Power calculations*



Randomisation will be conducted at three levels: school, department and teacher. The above power curves are for each of the three experiments. They all use the following two assumptions which were obtained from an existing key stage 3 to GCSE value-added model: intra-cluster correlation of 0.075 and correlation between key stage 3 and GCSE of 0.75. They make the following additional assumptions:

- n (intervention) =55; n (control) =65 represents the school randomisation and assumes an average cluster size of 180 (average cohort size for eligible secondary schools in England).
- n (high dose) =22; n (low dose) =23 represents the departmental randomisation and again assumes an average cluster size of 180 (average cohort size for eligible secondary schools in England).
- n (observer) =110; n (observee) =110 represents the teacher randomisation and will also have n (both) =110. This assumes an average cluster size of 30 for the size of each teacher's key stage 4 classes. Furthermore, it assumes an average of six teachers per subject per school (i.e. in an average school, a teacher would have one year 10 class and one year 11 class in their subject) and therefore a total of 110 teachers in each experimental group.

From the power curves, it can be seen the main trial is well powered with a minimum detectable effect size (MDES, at 80% power) of 0.1. The dosage trial has an MDES of around 0.15. Since we might expect the difference between intervention and control to be larger than between dosages we may miss a dosage effect if the main effect size is as low as 0.1. However, if the main trial demonstrates a larger effect, it may be possible to discern a dosage effect also. The teacher experiment appears very well powered with an MDES of 0.08 and this is absolutely necessary since we are not expecting differences between observer/observee/both to be as high as the effect of the main trial. In a scenario of low recruitment rates, the following recruitment thresholds have been calculated using the reasoning below.

| **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 |

Using the same assumptions as above, the main trial could be run with 25+25 schools and still be able to detect an effect of 0.15 with 80% power. However, this would render the dosage experiment underpowered. The teacher randomisation (observer/observee/both) would potentially still be viable here (50+50+50 teachers still able to detect an effect size of 0.12 with 80% power) but there would be a concern that the teacher randomisation is diluting the main effect and it would be better to have all teachers as both observers and observees if MDES=0.15. We should therefore only embark on the teacher randomisation when MDES=0.125 for the main trial (35+35 schools). It is not until we get to 50+50 schools that the dosage experiment can detect an effect of 0.15 with 80% power.

Since there is negligible cost associated with obtaining NPD data, all students in the trial can be included in the analysis of GCSE results. However, there is considerable cost associated with developing, printing, administering and marking the year 10 tests. For this reason, year 10 pupils will be randomly allocated to sit either the maths or English test in such a way that within each class, half the students will sit a maths test and half will sit an English test. Tests will be administered within schools by teachers. Despatch and collection of the tests will be managed by NFER.

### *Analysis*

The primary outcome will be mathematics and English GCSE outcomes[4]. Sub-group analysis on the primary outcome will be carried out on the following groups only: gender and whether or not a pupil receives free school meals (FSM). The secondary outcomes will be the two year 10 tests.

The main analysis will be 'intention to treat', reflecting the reality of how interventions are delivered in practice and avoiding attrition bias. We will use multi-level models with three levels (school, teacher, and pupil) to enable us to account for clustering and carry out analysis at the various levels of randomisation. Key stage 3 teacher assessment data will be used as a covariate in all models as this tends to have a higher correlation with GCSE performance than Key Stage 2 tests. Other variables that potentially explain some progress variance will be included in the models: gender, FSM, SEN, EAL and school type.

It is likely that the number of observations in the high and low dosage group will not be exactly as planned. Due to the method of data collection on observations (tablets), it should be possible to discern the exact number of observations carried out per teacher (both as observer and observee). The main analysis will therefore be followed by an 'on-treatment' analysis where data from the tablets will be used to determine the extent of each teacher's involvement. This analysis would enable us to estimate a 'pure intervention effect' (net of any fidelity issues) that would not necessarily be causal in nature.

### *Process evaluation*

The process evaluation will involve four main methods:

[4] There are two primary outcomes since the dosage randomisation is at the level of department and combining exam scores into a single outcome would not allow the effect of dosage to be discerned.

Firstly, we will carry out an online survey of all teachers of maths and English in both treatment and control schools, in the summer term of the first year. This will assess fidelity of the intervention in treatment schools and explore teachers' perceptions about the sustainability of the observation programme, different dosages and any practical issues concerning the use of the tablets or software. The survey in control schools will be to understand whether any compensatory behaviour occurred or whether they have alternative peer-observation programmes in place. Value for money analysis will also be included at this stage as we will ask teachers about the costs of the project to them including staff time and resources.

Secondly, NFER will visit 6 schools to interview teachers and senior management to understand implementation, school perceptions, barriers and necessary conditions for success in more detail. These visits will take place in spring 2015 and spring 2016. This will also allow the teachers to show how they are using the software and any implementation issues.

Thirdly, we will attend a training session and we will carry out a review of the training materials (we will also ask teachers about training in the survey). This will be in early spring term 2015.

Finally, we will carry out a review of records on observations and usage. This will take place in summer 2015 and summer 2016 after each year has been completed.

This combination of methods will give us a full understanding of how and why the intervention has/ has not worked including implementation challenges and adaptations, any unexpected outcomes and perceived impacts, perceived importance of, benefits from and commitment to the programme and participants views on its sustainability and suitability for national roll out.

At the end of the project, we will contribute to a post-analysis conference to feed back study findings to study participants. This will enable participants to help interpret and validate the study findings, identify any further points (for example, issues or impacts they do not feel have been adequately captured in the impact evaluation) and investigate the implications for roll out

Our reports in 2015 and 2016 will include the findings of the process evaluation and make recommendations to ensure the sustainability and replicability of teacher observation if it were scaled up.

## Personnel

The project will be led by Professor Simon Burgess/CMPO. The impact evaluation will be led by Dr Ben Styles and Dr Anneka Dawson at NFER. The process evaluation and Year 10 testing will be led by Juliet Sizmur at NFER. Camilla Nevill will have overview of the evaluation at EEF and Andy Cawthera will oversee the grant.

## Roles and responsibilities

Each person will carry out their duties with the assistance of teams at their respective institutions:

Simon Burgess – Recruitment and retention of schools, training and delivery of intervention, supply of list of eligible schools for randomisation

Ben Styles/ Anneka Dawson – trial management, randomisation, administration of tests and analysis.

Juliet Sizmur – process evaluation telephone interviews, visits, development and delivery of Year 10 tests.

## Data protection statement

NFER's data protection policy is available at:

http://www.nfer.ac.uk/nfer/about-nfer/code-of-practice/nfercop.pdf

| **Timeline** |
|---|
| |
| **Nov 2013:** Meeting with partner organisations, write and register protocol |
| **Jan-Feb 2014:** Pilot |
| **Apr- Jul 2014:** Recruit and consent schools and learners for main trial |
| **Aug 2014:** Randomisation of schools |
| **Sep- Oct 2014:** Training of teachers, attending a training session and review of video training materials |
| **Oct 2014 - July 2015:** Implementation of intervention programmes |
| **Apr 2015:** case study visits |
| **June 2015:** Process evaluation survey and GCSEs and Year 10 tests |
| **Aug- Sep 2015:** Review of usage data and reporting |
| **Oct 2015:** Interim report- including results from year 1 process and impact evaluations (without NPD) |
| **Sep 2015-July 2016:** Implementation of intervention programmes |
| **Jan 2016:** NPD data available for cohort 1 |
| **Apr 2016:** Case study visits |
| **June 2016:** GCSEs and Year 10 tests |
| **Aug- Sep 2016**: Review of usage data and reporting |
| **Jan 2017:** NPD data available for cohort 2 |
| **Feb- Mar 2017:** Post analysis conference |
| **Apr 2017:** Final report to EEF |

## Reporting

NFER will provide quarterly progress reports to EEF. After the pilot, we will provide a short summary report containing recommendations of any methodological changes for the main trial. We will provide a report for the EEF Board at the end of the first year of the evaluation (September 2015), reporting on data from the process evaluation and progress with the impact evaluation. We will prepare a final report of the evaluation findings to CONSORT standards in April 2017, which will include the findings of the impact evaluation together with the findings from the process evaluation. We will be pleased to discuss dissemination options, such as an article in a refereed or practitioner journal.

## Risks

| Risk | Assessment | Countermeasures and contingencies |
|---|---|---|
| School or teacher attrition | **Likelihood:** moderate<br>**Impact:** moderate | Clear information / initial meeting with schools explaining the principles of the trial and expectations. Both 'intention to treat' and 'on-treatment' analysis will be used.<br>Attrition will be monitored and reported according to CONSORT guidelines. |
| Interventions are not implemented well | **Likelihood:** low<br>**Impact:** moderate | Clear information / initial meeting with schools explaining the principles of the trial and expectations. Both 'intention to treat' and 'on-treatment' analysis will be used.<br>Process evaluation will monitor this. |
| Control schools increase level of their own teacher observation | **Likelihood:** moderate<br>**Impact:** moderate | Clear information / initial meeting with schools explaining the principles of the trial and expectations. Both 'intention to treat' and 'on-treatment' analysis will be used. |
| Delays in training of teachers and commencing interventions | **Likelihood:** moderate<br>**Impact:** low | Agree a clear timetable with project teams up front<br>Revise timetable for pre and post testing periods |
| Failure in recruiting schools | **Likelihood:** low<br>**Impact:** high | Project teams could make use of NFER's Research Operations Department to recruit more schools (at additional cost)<br>Timescale could be revised |
| Researchers lost to project due to sickness or absence | **Likelihood:** moderate<br>**Impact:** low | NFER has a large research department with numerous researchers experienced in evaluation who could be redeployed.<br>Senior staff can stand in if necessary. |
| Project teams do not follow correct trial protocols | **Likelihood:** moderate<br>**Impact:** high | Meetings with project teams at start of project.<br>Provision of clear guidance describing protocols for distribution to all schools. |