



E
E
F
ducation
ndowment
oundation

Teacher Observation

Evaluation report and executive summary
November 2017

Independent evaluators:

Jack Worth, Juliet Sizmur, Matthew Walker, Sally Bradshaw, Ben Styles



**Evidence for
Excellence in
Education**



Education Endowment Foundation

The Education Endowment Foundation (EEF) is an independent grant-making charity dedicated to breaking the link between family income and educational achievement, ensuring that children from all backgrounds can fulfil their potential and make the most of their talents.

The EEF aims to raise the attainment of children facing disadvantage by:

- identifying promising educational innovations that address the needs of disadvantaged children in primary and secondary schools in England;
- evaluating these innovations to extend and secure the evidence on what works and can be made to work at scale; and
- encouraging schools, government, charities, and others to apply evidence and adopt innovations found to be effective.

The EEF was established in 2011 by the Sutton Trust as lead charity in partnership with Impetus Trust (now part of Impetus - Private Equity Foundation) and received a founding £125m grant from the Department for Education.

Together, the EEF and Sutton Trust are the government-designated What Works Centre for improving education outcomes for school-aged children.



For more information about the EEF or this report please contact:

Danielle Mason

Head of Research and Publications
Education Endowment Foundation
9th Floor, Millbank Tower
21–24 Millbank
SW1P 4QP

p: 020 7802 1679

e: danielle.mason@eefoundation.org.uk

w: www.educationendowmentfoundation.org.uk

About the evaluator

The project was independently evaluated by a team from the National Foundation for Educational Research (NFER). The evaluation was directed by Dr Ben Styles and the team was led by Dr Anneka Dawson and Jack Worth. They were supported by Juliet Sizmur, Matthew Walker, Sally Bradshaw, and Sofia Farid.

Contact details:

National Foundation for Educational Research (NFER)

The Mere
Upton Park
Slough
Berkshire
SL1 2DQ

p: 01753 574123

e: j.worth@nfer.ac.uk

Contents

Executive summary	3
Introduction	5
Methods	11
Impact evaluation	20
Process evaluation	36
Conclusion	55
References	58
Appendix A: EEF cost rating	60
Appendix B: Security classification of trial findings	61
Appendix C: Observation domains and components	62
Appendix D: Feasibility study findings	65
Appendix E: Memorandum of Understanding	70
Appendix F: Randomisation SPSS syntax	77
Appendix G: Statistical Analysis Plan	83

Executive summary

The project

The Teacher Observation intervention aimed to improve teacher effectiveness through structured peer observation. Teachers observe and are observed by their peers a number of times over the course of two years. It was delivered by the Centre for Market and Public Organisation (CMPO) at the University of Bristol. CMPO researchers trained lead teachers from both maths and English departments in participating secondary schools to use RANDA TOWER software (RANDA, 2012), and lead teachers then trained colleagues in their schools. Teachers used the software on a tablet computer to keep a record of classroom observations and to review and collate the data afterwards. Intervention schools were requested to involve all maths and English teachers in a series of 20 minute, structured peer observations over a two-year period.

This study was an efficacy trial of the Teacher Observation intervention in 82 secondary schools with high proportions of pupils eligible for free school meals (FSM). It was a school-randomised controlled trial designed to determine the impact of the whole-school intervention on the attainment of pupils in English and maths. The primary outcome was the combined English and maths GCSE scores of Year 10 pupils in participating schools in 2014/15, who had two years of exposure to the intervention.

The trial was also designed to explore the effect of varying the number of observations undertaken by teachers ('dosage') and explore how the effect varied depending on whether the teacher was observing his or her peers or being observed or both. Teachers in the low dosage group were asked to complete a minimum of three observations (though the suggested number was six), while teachers in the high dosage group were asked to complete a minimum of four (with a suggested number of 12). A process evaluation used case study visits, an online survey and a pro forma to provide data on implementation and to capture the perceptions and experiences of participating teachers. The intervention was implemented in schools during the 2014/2015 and 2015/2016 academic years.

Key conclusions

1. The project found no evidence that Teacher Observation improves combined GCSE English and maths scores.
2. The project found no evidence of impact of the intervention on the GCSE English and maths attainment of pupils who have ever been eligible for FSM.
3. In general, teachers delivered the minimum number of observations allowed rather than the higher number suggested by the developers: teachers had difficulty fitting in the required number of observations because of timetabling and arranging cover, and some experienced problems using the software. Even when observations did take place, there was no evidence that schools which did more observations had better pupil results.
4. Teacher engagement with the programme varied greatly across schools, and practice ranged from individuals simply recording some observations using the RANDA software to whole-school, collaborative planning, discussion and reflection as part of an integrated CPD programme.
5. Almost three-quarters of the control group schools were already doing some peer observation prior to the intervention. The lack of impact seen in this study may be because the structured Teacher Observation intervention was no more effective than existing practice rather than because general peer observation has no impact.

EEF security rating

The findings from this study have very high security. The trial was a two-armed school-randomised controlled trial, which had a large number of schools and tested whether the intervention can work under

developer-led conditions. The final analysis had sufficient statistical power to detect an effect size of 0.11. Despite several schools dropping out of participation at various points during the evaluation, the final analysis included primary outcome data for pupils from all 82 randomised schools. Less than 10% of participating pupils had missing data, most of which was because of missing prior attainment (Key Stage 2) data. This attrition can be regarded as completely unbiased because it occurred before the schools were randomised. The intervention schools had similar pre-test pupil outcomes and numbers of FSM eligible pupils to the comparison schools.

Additional findings

The evidence from this evaluation suggests that the Teacher Observation intervention had no impact on pupils' GCSE English and maths attainment, compared to that of pupils in control schools. The analysis also suggests the intervention had no impact on FSM-eligible pupils' GCSE English and maths attainment. There was no evidence of an impact on pupils' GCSE English and maths attainment in departments selected to receive a higher dosage of observations compared to low-dosage departments. There was also no evidence of a differential impact on pupils' GCSE English and maths attainment depending on whether their teacher was the observer or was the one being observed teaching.



Implementation data suggests the number of observations was below the developer's initial expectations. The process evaluation found that teachers had difficulty fitting in the suggested number of observations because of timetabling and arranging cover, and some experienced problems using the software. However, our on-treatment analysis, which explored the association between the number of observations a teacher completed and the GCSE English and maths attainment of the pupils they taught, found no significant relationship.

Previous research suggests that pairing teacher observation with ongoing, school-based professional development is important for successful implementation. The process evaluation found that while some schools adopted the RANDA TOWER observation schedule as part of their ongoing CPD programmes and scheduled a number of additional planning, feedback and reflection sessions, in other schools, teachers conducted the observations but made no formal use of the materials beyond that.

Cost

Teacher Observation cost each school around £4000 per year, or £3 per pupil per year when averaged over 3 years. It also required several days of staff time for training and lesson observations. Participating schools also required iPads for the teachers to use the programme software.

Table 1: Summary of impact on combined English and maths GCSE scores

Group	Pupil numbers	Effect size (95% confidence interval)	p-value	Estimated months' progress	EEF security rating	EEF cost rating
Treatment vs control	7,366	-0.01 (-0.08, 0.06)	0.80	0		£££££
Treatment FSM vs. control	2,992	0.01 (-0.08, 0.10)	0.77	0		£££££

Introduction

Intervention

The Teacher Observation intervention aimed to improve teacher effectiveness through teachers observing their peers teaching and by being observed teaching by their peers a large number of times over the course of two years. The programme ultimately intended to improve learners' educational outcomes as a result of improvements in teacher effectiveness. Teachers in maths and English departments in intervention secondary schools used a tablet computer with RANDA TOWER software (RANDA, 2012) to record their observations. RANDA TOWER software allows teachers or leaders to use a tablet computer to record a number of classroom observations and, if required, review, collate or reflect on the data afterwards. The Teacher Observation intervention was delivered by the Centre for Market and Public Organisation (CMPO).

CMPO researchers trained lead teachers from both maths and English Departments in participating schools to use the RANDA TOWER observation software. The main aims and objectives of the training were:

- to introduce teachers to the project, project team, and the trial design;
- to explain the rubric/framework, and compare it with other competence models;
- to explore how the rubrics would be applied, to discuss descriptors and clarify outcomes; and
- to establish log-ons to the RANDA TOWER website, install the TOWER app on tablets, and explore the observation software.

Trained lead teachers then cascaded the training to colleagues within their own schools. Intervention schools were requested to involve all teachers in their maths and English departments in a series of 20-minute, structured peer observations over a two-year period. Teachers of Years 10 and 11 students in each department conducted either a higher or lower number of structured observations per annum using the RANDA TOWER software.¹

Teacher peer observations were focused on two domains. Within each domain, five component areas were identified for observation as summarised in Table 2. The detailed observation rubric provided for teachers' use is shown in Appendix C.

Table 2: Observation domains and components

Domain	Observed components
Classroom environment	Components which facilitate learning: <ul style="list-style-type: none"> • creating an environment of respect and rapport • establishing a culture for learning • managing classroom procedures • managing student behaviour • organising physical space

¹ The minimum number of observations was set at three per observer in the low dosage departments and four in high dosage departments over a two-year period. CMPO expected that the difference between the dosage categories would be larger than this; initially it recommended six observations per observer per annum for low dosage departments and twelve for high dosage departments

Teaching—skills which secure learning	<p>Components which secure learning:</p> <ul style="list-style-type: none"> • communicating with students • using questioning and discussion techniques • engaging students in learning • use of assessment • demonstrating flexibility and responsiveness
---------------------------------------	---

During their 20-minute observations, teacher observers rated each component according to four levels of competence:

- ineffective (1–3);
- basic (4–6);
- effective (7–9); or
- highly effective (10–12).

Teachers in each department were divided into three groups by random allocation: one third only conducted observations ('observers'), one third were only observed ('observees'), and one third both observed and were observed ('both'). Beyond the number of observations expected of each group, no further requirements were specified by CMPO, for example, schools were free to plan observations and/or review and use the observation data (or not) as they wished.

Participants were encouraged to use the information and support pages on the RANDA TOWER website (which only licensees can access) and to contact CMPO with queries at any time throughout the project. CMPO monitored the number of observations conducted by the intervention schools and provided support as necessary to encourage completion of the targeted number.

There was no requirement for post-observation discussion to take place as part of the intervention, although lead teachers were prompted to consider using the outcomes of observation to focus a post-observation discussion during the training.

Background evidence

Research evidence suggests the most important action that schools can take to improve outcomes for students is supporting their teachers to be more effective (Mourshed *et al.*, 2010). The research literature suggests that the most reliable way to achieve this is to develop a professional culture where teachers are continually adapting and refining their skills and methods (Cordingley *et al.*, 2015). Yet, while improving the quality of teaching is critical to student success, it is only recently that educational administrators have begun to take seriously the importance of systematically evaluating teacher performance.

One approach to teacher evaluation is to use highly-structured classroom observations as a means of providing teachers with the feedback they need to improve. In England, classroom observation has a long-standing role in the pre-service preparation and continuing professional development (CPD) of teachers. Here, it has largely been used formatively to provide feedback on performance or to model alternative teaching approaches (O'Leary and Brooks, 2013). In the last two decades, however, it has been increasingly appropriated as a policy tool that seeks to combine its original formative purpose with a new focus on accountability (O'Leary and Brooks, 2013; Richards, 2014).

Research from Australia suggests that no one understands the importance of introducing more effective teacher evaluation systems better than teachers themselves. Analysis of OECD TALIS² data found that 63% of teachers report that appraisal of their work is largely done to fulfil administrative requirements, while 61% report that appraisal of their work has little impact on the way they teach (Jensen and Reichl, 2011). For many teachers, therefore, teacher evaluation is largely seen as a bureaucratic exercise that is not linked to teacher development or improved classroom teaching.

In response to these challenges, new teacher evaluation systems have emerged, particularly in the U.S. Fuelled by incentives from the federal government, U.S. state and local policymakers have sought to replace the often cursory evaluation models of the past with more comprehensive ones (White, 2014). The Teacher Advancement Program (TAP) is one such model. Originally developed in the late 1990s, it was introduced to Chicago in 2007 as a school-wide reform model (Glazerman and Seifullah, 2012). Under the Chicago TAP model, the school leadership team undertake regular classroom observations. The idea behind TAP was that by giving teachers performance incentives, along with tools to track their performance and improve instruction, schools would attract and retain talented teachers and help all teachers raise student achievement. However, the findings from a four-year evaluation (2007–2011) found that TAP was only partially successful in achieving its goals. While Glazerman and Seifullah (2012) found evidence that Chicago TAP had increased schools' retention of teachers, it did not have a noticeable positive impact on student achievement. White (2014) argues that Chicago's teacher evaluation system highlights some of the challenges of observing teaching, and that such systems cannot rely solely on school principals to conduct the classroom observations as the time demands present a substantial burden.

In an evolution of this approach, 'multi-rater' systems, whereby more than one observer rates a teacher, have been proposed as a possible solution (White, 2014). In the largest study of instructional practice ever undertaken, the Bill and Melinda Gates Foundation's Measures of Effective Teaching (MET) project set out to investigate how a set of measures could identify effective teaching fairly and reliably. With the help of 3,000 teacher volunteers, their analyses revealed that adding a second observer increases reliability significantly more than having the same observer score an additional lesson. They also found that additional, shorter observations can increase reliability, that school administrators rate their own teachers somewhat higher than do outside observers, and that adding observations from outside a teacher's school to those carried out by a teacher's own administrator can provide an ongoing check for in-school bias (Bill and Melinda Gates Foundation, 2013). Similar findings have been found in other studies. In a study of a 'dual-rater' classroom observation scheme in 20 elementary and middle schools in the Western U.S., Manzeske *et al.* (2014) found that principals rated their staff more positively than outside observers. They also found that principals exhibited a greater range of variability in their scoring, demonstrating less consistency than outside observers. MET researchers also found that while there is no hard-and-fast requirement governing the number of observations or the number of observers, ratings became more reliable with each individual observation, so anything more than a single classroom visit per year offers an incremental improvement (TNTP, 2012).

Among the observational protocols currently available, one of the most commonly used in the U.S. is the Danielson Framework for Teaching (FfT). The FfT rates teachers at four levels of performance—unsatisfactory, basic, proficient, or distinguished—and includes four domains of teacher effectiveness. These capture planning and preparation, classroom environment, instruction, and professional responsibilities (Garrett and Steinberg, 2015). The FfT was used as part of both the Chicago TAP and MET projects.

In addition to the examples already discussed, the research literature provides examples of where observation-based teacher evaluations, combined with effective professional development, have impacted positively on student outcomes. A study by Shaha *et al.* (2015) based on 292 schools within

² OECD Teaching and Learning International Survey: a survey of lower secondary teachers and their school leaders around the world. <http://www.oecd.org/edu/school/talis.htm>.

27 U.S. states found that teacher observations, coupled with online, on-demand professional development, resulted in significantly improved student achievement in reading and maths on standardised assessments. For example, schools with Lower Observation Rates (with a mean number of teacher observations of 2.76) collectively experienced 2.22 net gain in the percentage of students rated as ‘Proficient’ or ‘Advanced’ representing a statistically significant 3.9% improvement. In contrast, schools with Higher Observation Rates (with a mean number of teacher observations of 8.49) collectively experienced 13.16 net gain in the percentage of students classified as ‘Proficient’ or ‘Advanced’ representing a statistically significant 24.9% improvement from baseline. The comparison reflects 6.37 times the growth in the Higher Observation Rate schools compared with the Lower Observation counterparts. Similarly, a study of mid-career elementary and middle school teachers in Cincinnati Public Schools found that teachers were more effective at raising student achievement during the school year when they were being evaluated than they were previously (Taylor and Tyler, 2012).³ Participating teachers were evaluated over a year, based largely on classroom observation. The effect on student attainment was relatively small, at 0.11 standard deviations for maths attainment, but had not been tested using a randomised controlled trial (RCT). It was anticipated that the impact of the structured observation would continue to develop over time, particularly if schools adopted a cultural shift towards regular peer observation. Collectively, these findings suggest that effective teacher evaluation systems based on well-structured teacher observation can enhance teacher effectiveness and raise student attainment.

Against this background, the rationale for setting up the Teacher Observation intervention was to explore whether teacher observation does improve teacher effectiveness.⁴ The intervention is based on observation, not evaluation, and does not attach incentives or penalties to performance. The structure for the observation came from the well-tested rubric from the U.S. and was adapted by CMPO for the English context to incorporate features of the current Teachers’ Standards (DfE, 2011) and OFSTED descriptors (Ofsted, 2016). The final rubric is provided in Appendix C.

A small feasibility study was conducted by CMPO and NFER in 2014–2015. The aims of the feasibility study were to:

- test how the RANDA software and rubric worked and identify/solve any technical issues for delivery;
- gain an understanding of the technical requirements for the software, such as broadband;
- carry out a small-scale process evaluation involving case study visits to the schools to interview senior leaders and those taking part in the feasibility study; and to
- trial/pilot the Year 10 tests with a view to exploring and refining the reliability and validity of the assessment instruments to be used in the main trial.

After positive feedback on the functionality of the software and some U.K.-specific amendments to the rubrics, intervention was deemed ready for an efficacy trial. The feasibility study also provided trial item data for the development of the bespoke Year 10 tests. NFER’s findings from the feasibility study are shown in Appendix D.

Evaluation objectives

The primary research question for the evaluation was:

- What is the impact of two years of the Teacher Observation programme on learners’ GCSE maths and English achievement?

³ The rationale and processes for the current Teacher Observation study were based on the work of Taylor and Tyler, 2012.

⁴ That is, a teacher’s ability to raise student achievement.

The secondary research questions were:

- What is the impact of Teacher Observation on learners' Year 10 mathematics and English achievement?
- What is the impact of one year of Teacher Observation on learners' GCSE maths and English achievement?
- What effect does varying the minimum expected number of observations have on learners' pupil attainment?
- Does the impact of the intervention on pupil attainment vary depending on whether a teacher is an observer, observee, or both?

The objectives of the process evaluation were to:

- assess fidelity of the intervention in treatment schools and explore teachers' perceptions about the sustainability of the observation programme using an online survey of teachers in intervention schools;
- understand implementation, participants' views on its sustainability and suitability for national roll out, and barriers and necessary conditions for success in more detail by visiting six schools to interview teachers and senior management;
- assess the level of implementation by reviewing records on observations and usage; and
- understand whether any compensatory behaviour occurred in control schools, or whether they have alternative peer-observation programmes in place, with a survey of teachers in control schools.

The EEF published the evaluation protocol setting out the above objectives when the project began in July 2014. An amended protocol was published in October 2014 reflecting changes to the evaluation as a result of the pilot phase.⁵ The subsequent deviations from the protocol are highlighted in the report alongside an explanation of why the deviation occurred.

Ethical review

The evaluation and consent procedure was approved by CMPO's ethics board at the University of Bristol as well as by NFER's Code of Practice committee. CMPO obtained headteacher consent from all the schools taking part in the evaluation via a memorandum of understanding (MOU). The document explained the trial and the responsibilities of the school, CMPO, and NFER, and included a statement on how schools' data would be used throughout the evaluation (an example MOU is shown in Appendix E). When schools withdrew from the intervention or evaluation, we sought the school's explicit permission to match the pupil data we had previously collected to the National Pupil Database (NPD). Of the 45 schools who withdrew from participation in various aspects of the project (implementing the intervention, providing pupil-teacher linked data, or participating in testing of Year 10 pupils), 31 allowed permission to match the pupil data to the NPD, 13 refused permission, and one did not respond. We did not match pupil records for these 14 schools to the NPD, but did obtain anonymized pupil data from the NPD for these schools for our analysis.

In addition, we obtained consent from schools participating in the summer 2015 Year 10 tests to match their unique pupil reference numbers (UPNs) with the NPD, and to extract names and dates of birth which we used to pre-populate the Year 10 tests. Ten of the 70 schools we sent tests to did not consent to us matching their UPN data with the NPD, so we pre-populated their tests with the students' UPNs.

⁵ <https://educationendowmentfoundation.org.uk/our-work/projects/teacher-observation/>

One of the 46 schools we sent tests to for the summer 2016 Year 10 tests was pre-populated with the students' UPNs for the same reason.

Project team

The project was led by Professor Simon Burgess at the Centre for Market and Public Organisation (CMPO), University of Bristol. Simon was responsible for the design of the evaluation, recruitment and retention of schools, and training and delivery of the intervention.

The impact evaluation was directed by Dr Ben Styles at NFER. Dr Anneka Dawson managed the trial for the first eighteen months. This role was then handed to Jack Worth who was the trial manager for the remainder of the trial. He was assisted by Sally Bradshaw on the statistical analysis and Sofia Farid on school liaison and operations management. The process evaluation and Year 10 testing were led by Juliet Sizmur with assistance from Matt Walker.

Trial registration

The trial was registered on the ISRCTN registry on 23 June 2014.⁶

⁶ <http://www.isrctn.com/ISRCTN89620259>

Methods

Trial design

The trial was designed as a set of three randomised experiments within the same schools, with the randomisation occurring at different levels.

Experiment 1 —school-level experiment

Experiment 1 was a school-randomised controlled trial of the observation intervention. The school-level experiment was the main experiment, testing whether the intervention had an impact on pupil outcomes. Schools were randomly assigned to either Teacher Observation (subsequently referred to as the ‘intervention group’) or a ‘business as usual’ control (‘control group’). Financial incentives, in the form of two £500 payments on completion of the Year 10 tests at the end of each academic year, were provided to control schools to encourage them to continue participating in the evaluation.

Experiment 2—department-level experiment

Experiment 2 was a department-level dosage randomisation within intervention schools. English and maths departments in each intervention school were randomly assigned to either a high or low dosage of observations so that each school had one low observation category and one high. The number of observations suggested for the dosage categories was six per year in the low observation category and 12 per year in the high observation category. However, the pilot revealed that it was not 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, it was agreed that the minimum number of observations would be set at three in the low dosage departments and four in high dosage departments. However, CMPO expected that the difference between the dosage categories would be larger than this.

Experiment 3—teacher-level experiment

Experiment 3 was a teacher-level randomisation, also only within intervention schools. All English and maths teachers in each intervention school were randomly assigned to one of three groups: observers (who only observed), observees (who were only observed by others) and a group that did both (they observed and were observed). NFER randomised a set of teacher IDs assigned to schools by CMPO so that only schools could identify each teacher.

The trial was designed by CMPO with some minimal input from NFER concerning sample size. The amended protocol lists changes to the original design, the main one being the definition of high and low dosage (see ‘Experiment 2’ above).

Participant selection

In summer 2014, CMPO contracted recruitment consultants to approach a sample of English state-funded secondary schools that were in the highest 50% of schools ranked by percentage of pupils eligible for FSM. Single sex schools and schools with boarders were excluded due to unobservable differences in these schools. At the EEF’s request, secondary schools in Lancashire, Merseyside, and Somerset were excluded from the sample due to ongoing recruitment for other EEF-funded projects; schools in all other local authorities in England were eligible for inclusion in the sample. In total, 1,097 schools were approached. CMPO sought headteacher consent from schools prior to randomisation via a memorandum of understanding (MOU). An example MOU is shown in Appendix E.

The initial plan was for CMPO to recruit 120 schools. However, by the beginning of the intervention delivery period in September 2014, 93 schools had been recruited and had returned signed MOUs. Of these, one Welsh school was excluded as it did not meet the EEF's eligibility criteria.⁷

To be included in the trial, schools needed to provide a list of unique pupil numbers (UPNs) for Year 10 and Year 11 pupils, and class lists linking anonymous teacher IDs to the UPN of pupils they taught. We put this in place to avoid control schools resisting the data requirements after allocation. While it is NFER's preference to request names and dates of birth with UPNs in advance of a trial to avoid mismatches with the NPD, CMPO felt that schools would be more willing to supply lists of UPNs without this information.

The impact of the intervention was not necessarily restricted to Year 10 and 11 pupils as teachers influenced by the intervention will have taught pupils from across the school. However, we focused on Year 10 and 11 pupils to estimate the impact of the intervention on attainment because their GCSE outcomes could be tracked in the National Pupil Database.

Outcomes measures

As an indicator of educational attainment, the primary outcome measure is GCSE examination results in English and mathematics—the subjects taught by intervention teachers. GCSE point scores (a standard measure that translates letter grades into a numerical score) for each pupil in both subjects were accessed from the NPD. Almost every pupil in secondary school sits GCSE examinations in both subjects, maximising the data available for analysis. Using data from the NPD ensured that we would have data for every pupil who began the trial in one of the randomised schools regardless of whether the school participated in the project to the end or whether a pupil moved school before sitting his or her GCSE examinations.

For the school-level experiment we used the sum of the two GCSE scores as the measure of pupil attainment. For the department- and teacher-level experiments we used the GCSE score for the subject taught by the particular department or teacher as the outcome variable.

We also administered Year 10 maths and English tests as a secondary outcome measure of attainment for Year 10 pupils. Although some commercial maths and English tests that cover basic skills are available, we did not feel that these were adequately tailored to provide a valid assessment of the types of changes in student performance that might be expected to result from this intervention. We therefore constructed new, bespoke tests by selecting items from Key Stage 3 past papers covering the ten-year period from 2000 to 2010. We knew that these were robust test items as they were rigorously developed and trialled for the Qualifications and Curriculum Development Agency (QCDA) or the Standards and Testing Agency (STA). They also cover a wide range of ability and are appropriate to the age range and the curriculum. Our aim was to provide a valid measure that was more likely to detect subtle changes in student performance, compared with commercially available instruments.

Year 10 test development began in the spring and summer terms of 2014 when we conducted a series of expert reviews of all existing Key Stage 3 tests in maths and in English, categorising items according to the specific skills required and the level of challenge presented. We made an initial selection of twice the number of items we would need to ensure broad coverage of the curriculum in both areas. As the tests had to be administered within one hour, the total number of items included was necessarily restricted. We agreed, therefore, that the majority of items should be focused on higher order skills rather than those related to rote learning as these would be the types of thinking skills that should be

⁷ Ten schools withdrew from participation in the evaluation shortly afterwards, before they were informed of their allocation to intervention or control, so the final analysis includes 82 schools (see 'Participants' section below).

improved through this type of intervention.⁸ So, in the reading tests, questions focused more on inference and interpretation than on simple retrieval of information, and in maths, questions were selected that allowed students to demonstrate their understanding and application of mathematical concepts and included elements of analysis and evaluation as opposed to simple remembering. We also trialled some items informally with students before finalisation for the test pilot. The maths tests consisted entirely of published Key Stage 3 items. The English tests included one unit developed by NFER English test development experts to include an element of English writing, something not present in Key Stage 3 English tests.

In June and September 2014, we piloted two maths and two English tests in schools with matching demographics to those being recruited for the main intervention study. The purpose of the pilot was to ensure that the final Year 10 tests were of an appropriate level of difficulty and provided maximum discrimination between students performing at different levels.⁹ A team of NFER specialist markers then scored around 400 student responses for each before conducting a full item analysis. The pilot test statistics indicated that all four tests showed high reliability (that is, they were addressing coherent underlying constructs in maths and English). However, some of the trial maths items proved to be too challenging for many of the Year 10 students in schools of the type recruited for the project. Test items found to be too hard (or too easy) were discarded from the pool and final tests were constructed to ensure the most thorough coverage possible within a one-hour test.

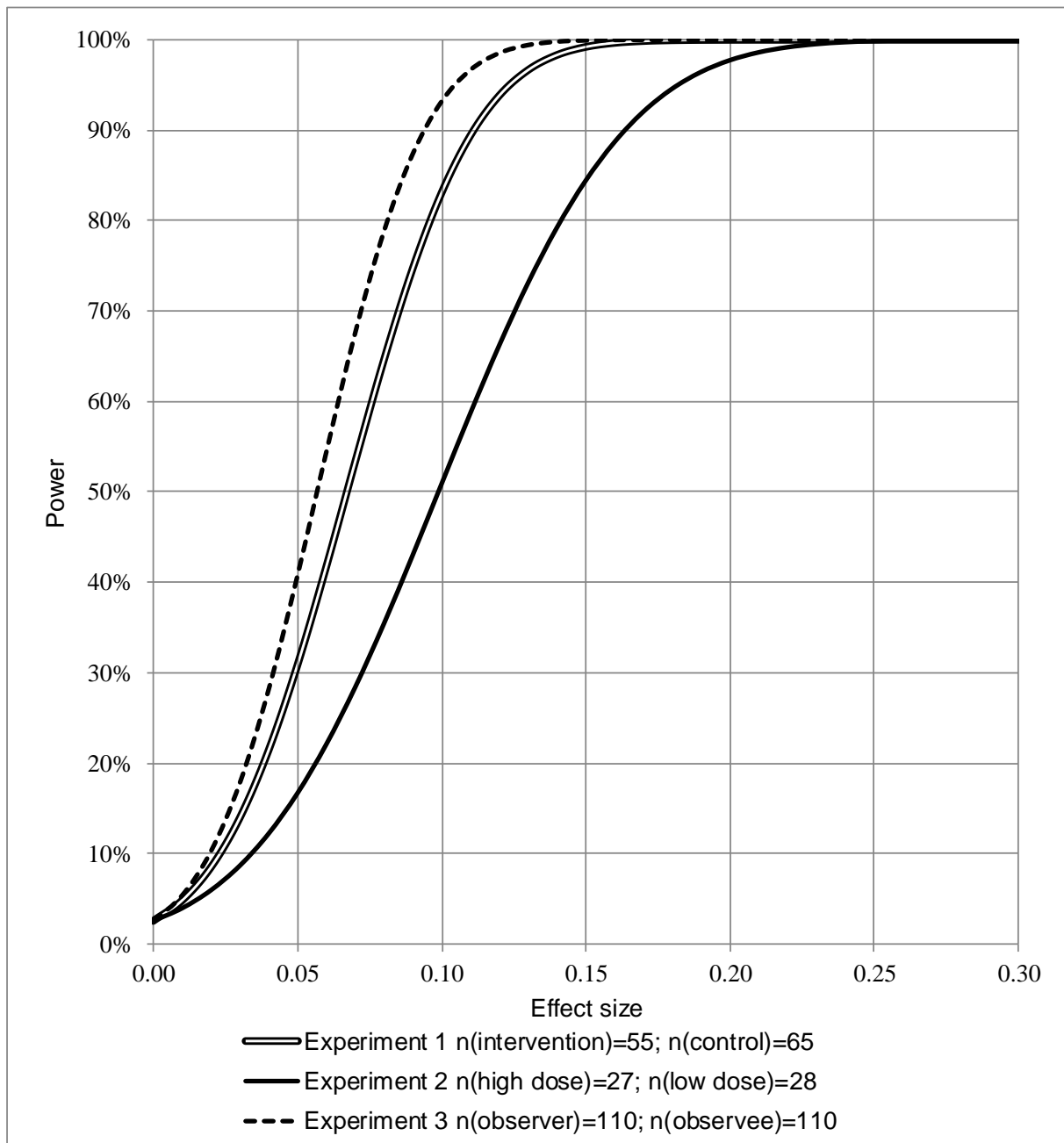
For the final tests, Year 10 students were randomly allocated to sit either the maths or the English test in such a way that, within each class, half the students sat a maths test and half an English test. This was due to the considerable cost associated with printing, administering, and marking all the tests. Teachers administered the tests in June and July 2015 and June and July 2016 under exam conditions. NFER managed the despatch and collection of the tests and they were scored by NFER markers. The markers did not know which tests had been taken by pupils in the intervention or the control group (scoring was blinded).

Sample size

We conducted randomisation at three levels: school, department, and teacher. The power curves in Figure 1 summarise our anticipation of the power of each of the three experiments at the project outset.

⁸ The aim of the teacher observation project was to support teachers in developing (what they believed to be) most effective teaching practice and, thereby, improve overall student achievement. It is generally held that high achieving students demonstrate more highly developed skills of analysis, synthesis, and evaluation rather than simply memorising (knowing), understanding, and applying.

⁹ Standardisation data is not essential where both intervention and control groups are being tested on the same instruments.

Figure 1: Power curves for the three experiments

All three power calculations use the following two assumptions that 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 0.75.¹⁰ They make the following additional assumptions:

- **Experiment 1 (school-level experiment):** $n(\text{intervention}) = 55$; $n(\text{control}) = 65$ represents the school randomisation and assumes an average cluster size of 180 (average cohort size for eligible secondary schools in England).

¹⁰ The final analysis uses Key Stage 2 scores as a covariate rather than Key Stage 3 because Key Stage 3 tests are no longer compulsory and data is no longer available from the NPD.

- **Experiment 2 (department-level experiment):** n (high dose) = 27; n (low dose) = 28 represents the departmental randomisation and again assumes an average cluster size of 180 (average cohort size for eligible secondary schools in England).¹¹
- **Experiment 3 (teacher-level experiment):** n (observer) = 110; n (observee) = 110 represents the teacher randomisation and will also encompass n (Both) = 110. This assumes an average cluster size of 30 pupils in each teacher's Key Stage 4 classes and an average of six teachers per subject in each of the 55 intervention schools.

Based on these assumptions, the school-level experiment has a minimum detectable effect size (MDES, at 80% power) of 0.1. The department-level experiment has an MDES of around 0.15. The teacher-level experiment has an MDES of 0.08.

We defined thresholds for low-recruitment scenarios (Table 3) to ensure each experiment had sufficient power, acknowledging the intervention dilution effect of the other experiments (for example, introducing dosage differences within the intervention schools dilutes the school-level intervention versus control comparison).

Table 3: Low recruitment thresholds.

Number of schools	Trial design
50 or more	Proceed with school-level experiment only (all departments high-dosage and all teachers to be 'both' (i.e. observers and observees)).
70 or more	Proceed with school- and teacher-level experiment (all departments high-dosage).
100 or more	Proceed with all three experiments.

The final recruitment yielded 92 schools and CMPO and the EEF decided to proceed with all three experiments but on the basis that the departmental-level experiment would be considered exploratory. This was because it involved a less extreme comparison than the school-level experiment but used half the number of clusters.

Randomisation

An NFER statistician carried out randomisation at all levels using SPSS with a full syntax audit trail and a random number generator to assign groups. Syntax for the school- and department-level randomisations and example syntax of the teacher-level randomisation are shown in Appendix F.

Schools were allocated to groups (46 intervention and 46 control) using stratified randomisation to balance across a range of school-level measures. The stratifiers used, defined by CMPO, were:

- percentage FSM pupils—split into a binary variable above or below median in the sample;
- percentage of white ethnicity students—split into a binary variable above or below median in the sample; and
- school value-added score across maths and English—CMPO calculated value-added for each subject by looking at progress from Key Stage 2 to GCSE accounting for pupil gender, major ethnic group, and FSM; CMPO then calculated the mean across subjects and split this into a binary variable above or below the median in the sample.

The stratifiers were chosen as each of the above factors could influence outcomes. We used simple randomisation to randomly allocate departments within intervention schools to low- or high-dosage

¹¹ In the protocol the numbers in each group were 22 and 23. These should have read 27 and 28 as there were 55 intervention schools in Experiment 1.

categories in the ratio 1:1 to guarantee that each school had one high- and one low-dosage department. Teachers in intervention schools were randomly allocated to be observers, observees, or both in the ratio 1:1:1 by randomising their unique teacher ID using simple randomisation. Schools provided CMPO with the teacher IDs; neither CMPO nor NFER received any teacher name data. When new teachers joined the school during the year they were allocated a new teacher ID. If the new teacher was directly replacing a teacher that had left, they took on the same role as the teacher who had left (observer, observee or both). However, if the new teacher was additional to the current staff or was not a direct replacement then NFER randomly allocated them into a role. There were 15 extra randomisation events during the course of the trial to address this type of scenario.

Analysis

We have conducted the analysis in line with the EEF's analysis guidance. We used multi-level modelling to estimate the impact of the intervention on outcomes because each experiment was cluster-randomised.¹² School/department was the first level and pupil was the second level in the models for the school- and department-level experiments. School, teacher, and pupil levels were included in the analysis of the teacher-level experiment. We calculated the standardised effect size by dividing the coefficient of the intervention indicator from the multilevel model by the standard deviation of the outcome variable in the whole sample (intervention and control groups). This standard deviation was derived from the total variance of a multilevel model without covariates.

Many pupils in the dataset were taught by more than one teacher which meant that the data for analysing the teacher-level experiment needed structuring in a different way from the school-level experiments. Some pupils were taught by multiple teachers from the class lists provided by schools (shared classes) and many pupils had different teachers in the Year 10 class lists and the Year 11 class lists (changes of staffing). We restructured the pupil-level data from the NPD into a pupil-teacher-level dataset where each case in the dataset has a pupil ID and a teacher ID and data relating to that pupil and that teacher. Pupils and teachers have multiple records in the dataset. Following Slater *et al.* (2009), we ensured that the number of pupils was correctly counted in our analysis by weighting the data.¹³ The sum of weights for each pupil in each subject across both years of the evaluation equalled one. Where schools provided class list data in the first year but not the second year of the evaluation, the sum of weights for each pupil equalled 0.5. Table 4 illustrates some examples of how the data is structured and how weights have been applied.

Table 4: Examples of pupil-teacher data structure and weights

Pupil ID	Year of study	Subject	Teacher ID	Weight
Pupil 1	1	English	Teacher E1	0.5
Pupil 1	2	English	Teacher E2	0.5
Pupil 1	1	Maths	Teacher M1	0.5
Pupil 1	2	Maths	Teacher M2	0.5
Pupil 2	1	English	Teacher E3	0.25
Pupil 2	1	English	Teacher E4	0.25
Pupil 2	2	English	Teacher E3	0.25

¹² Teachers, who are individuals, were randomised for the teacher-level experiment but they are regarded as a cluster for analysis because they teach multiple pupils, which is the unit of analysis.

¹³ Correctly estimating the standard errors and confidence intervals depends on using the correct number of pupils. Without weighting, the statistical software would take the number of pupil-teacher links which would underestimate the standard error and the width of confidence intervals.

Pupil 2	2	English	Teacher E4	0.25
Pupil 2	1	Maths	Teacher M1	0.5
Pupil 2	2	Maths	Teacher M2	0.5
Pupil 3	1	English	Teacher E5	0.5
Pupil 3	1	Maths	Teacher M3	0.5

Note: Pupil 1 had different English and maths teachers for both years. Each pupil-teacher record has a weight of 0.5 to ensure the pupil weights sum to one for each subject over both years. Pupil 2 has the same two English teachers for both years. Each of these pupil-teacher records has a weight of 0.25 to ensure the pupil weights sum to one for English over both years. Pupil 3 has one English teacher and one maths teacher in the first year, but the school did not submit class lists in the second year. Therefore, the pupil weights sum to 0.5 for each subject to be consistent with other pupils.

Pupils' Key Stage 2 scores were used as a covariate in all the models. We used KS 2 reading where English (GCSE or Year 10 test) was the outcome variable and KS 2 maths where mathematics (GCSE or Year 10 test) was the outcome variable. The primary analysis for the school-level experiment used combined KS 2 scores to mirror the combined GCSE scores as the outcome variable. It also included the randomisation stratifiers as covariates. We extended the primary analysis of the school-level and department-level experiments to consider the impact on the sub-group of pupils who have ever been eligible for FSM. Where possible, the prior attainment covariate was the fine grade point score derived from KS 2 tests. However, the partial boycott of 2010 KS 2 testing meant that a large portion of fine grade data was missing for the 2015 GCSE cohort (Cohort 1, see Table 6). We used a teacher-assessment-based measure as a covariate for our secondary analysis of this cohort's outcomes. This resulted in a lower correlation between the pre-test and post-test measures and therefore less statistical precision.

We conducted all analysis as intention-to-treat, except for some on-treatment analysis that related the number of observations conducted by each teacher to the attainment of the pupils they taught. On inspection of the final data, we decided it was preferable to perform the on-treatment analysis using the pupil-teacher dataset rather than the pupil-level dataset (a change to the analysis proposed in the SAP). This was so that the number of observations that each teacher had conducted could be related to the pupils they taught rather than calculating a department-level average of the number of observations conducted. For this on-treatment analysis a measure of the number of observations replaced the intervention group indicator in the multilevel regression model. Further on-treatment analysis distinguished between the number of observations each teacher had been involved in as the observer and as the teacher being observed.

Less than 10% of pupil data was missing for the primary analysis of the school-level and department-level experiments. This anonymized pupil data was matched by the NPD team to our list of randomised schools rather than to the CMPO pupil lists obtained from the schools. The reason for most of the missing data was pupils not having KS 2 point scores to use as a prior attainment covariate (see participant flow diagrams below). As this data was missing before randomisation we can regard it as unbiased, so further analysis of missing data (such as multiple imputation analysis) is not warranted.

See Appendix G for the statistical analysis plan. This was developed and published in January 2017 and provides full details about how we conducted the analysis.

Implementation and process evaluation

The process evaluation involved the following methods:

- observing a CMPO-led training session and reviewing training materials to assess how teachers were prepared for the intervention;

- surveying teachers and senior leaders, in both intervention and control schools, to 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—in control schools, the survey also explored whether any compensatory behaviour occurred or whether they had alternative peer observation programmes in place;
- visiting six intervention case study schools in the first year to interview teachers and senior management to understand implementation, school perceptions, barriers, and necessary conditions for success in more detail—we selected case study schools to represent a wide range of school types, dosage, and geographical location;
- analysing RANDA usage data to evaluate fidelity to the intervention in terms of the number of observations carried out (high and low dosage);
- carrying out follow up telephone interviews with case study schools in the second year of the programme to ascertain whether any new issues had arisen; and
- sending e-mail pro formas incorporating focused, open-ended questions to identify the perceived benefits, drawbacks, and barriers to implementation at the end of the two-year project, and, where possible, to ascertain levels of engagement and costs.

NFER researchers collected all the process evaluation data. We reassured participants of their anonymity, ensuring that they felt comfortable to respond openly throughout.

It should be noted that although the original protocol listed six school visits in the second year of the intervention, we agreed with the EEF that we would instead send open-ended pro formas to co-ordinators in *all* intervention schools and undertake selected telephone interviews on the basis that this approach was more likely to provide answers to any outstanding issues (May 2016).

Costs

We gathered information about the cost of the intervention from CMPO and from teachers through the teacher survey and pro forma. This included the overall costs for software licences and training, additional time or financial costs borne by schools, and any pre-requisite costs. We established which costs were one-off and which were ongoing.

We estimated a per-school cost based on the number of schools that were in the evaluation. We then divided the cost per school by the typical number of pupils in a school (rather than the number of pupils that formed part of the evaluation) assuming that if a school implemented this intervention then it would affect all its pupils. We profiled the cost over three years by considering the one-off and ongoing costs separately, as per the EEF guidance.

Timeline

Table 5: Timeline

Date	Activity
Nov 2013	Meeting with partner organisations, write and register protocol
Jan–Feb 2014	Pilot
Apr–Jul 2014	Recruit and obtain consent from schools for main trial
Jun 2014–Feb 2015	Pilot Year 10 tests
Sep–Oct 2014	Randomisation of schools

Oct 2014	Training of teachers and attending a training session
Oct 2014–Jul 2015	Implementation of intervention programmes
Apr 2015	Case study visits
Jun–Jul 2015	Process evaluation survey; GCSEs and Year 10 tests
Aug–Sep 2015	Review of usage data and reporting
Oct 2015	Interim report to the EEF (unpublished) including results from Year 1 process and impact evaluations (without NPD)
Sep 2015–Jul 2016	Implementation of intervention programmes
Jan 2016	NPD data available for Cohort 1
Apr 2016	Pro formas and telephone interviews
June 2016	GCSEs and Year 10 tests
Aug–Sep 2016	Review of usage data and reporting
Jan 2017	NPD data available for Cohort 2
Apr 2017	Draft evaluation report to the EEF

Impact evaluation

Participants

As explained in the ‘participant selection’ section above, CMPO approached 1,097 schools with the highest percentages of students eligible for FSM. One of the 93 recruited schools was excluded because it was a Welsh school (not meeting the EEF’s eligibility criteria). NFER randomised the remaining 92 schools, but ten schools withdrew from participation in the evaluation before they were informed of their allocation to intervention or control. Five of the schools had been allocated to the intervention group and the other five had been allocated to the control group. We retained the randomisation allocations for all schools and regarded this withdrawal from the evaluation as unbiased. As the withdrawal of these ten schools is unbiased, the final analysis uses data from the 82 schools who were informed of their randomised allocation.

We measured the outcomes of three different cohorts of pupils at several points throughout the evaluation to measure the impact of the intervention on their attainment. Table 6 shows the definition of the different cohorts and what outcomes we measured for each cohort as part of the evaluation. The primary cohort, Cohort 2, had two years of exposure to the intervention; the primary outcome is the GCSE results of Cohort 2.

Table 6: Definition of the cohorts of pupils

Cohort	Definition	Outcomes measured in:	
		2014/2015	2015/2016
Cohort 1	Year 11 in 2014/2015	GCSE English & maths	
Cohort 2	Year 10 in 2014/2015	Year 10 tests	GCSE English & maths
Cohort 3	Year 10 in 2015/2016	Year 10 tests	

The participant flow diagrams for the school-level (Figure 2), department-level (Figure 3 for English and Figure 4 for maths), and teacher-level experiments (Figure 5) are shown below for the primary analysis of Cohort 2. Randomised and allocated pupils are defined as those who were Year 10 pupils on roll at an intervention or control school in the school census of autumn 2014. Pupils were lost to follow-up if they did not sit GCSE examinations. Very few pupils fell into this group because almost all schools enter almost all of their pupils for English and maths GCSEs and pupils’ school census records in the NPD are linked to their GCSE results even if they move school.

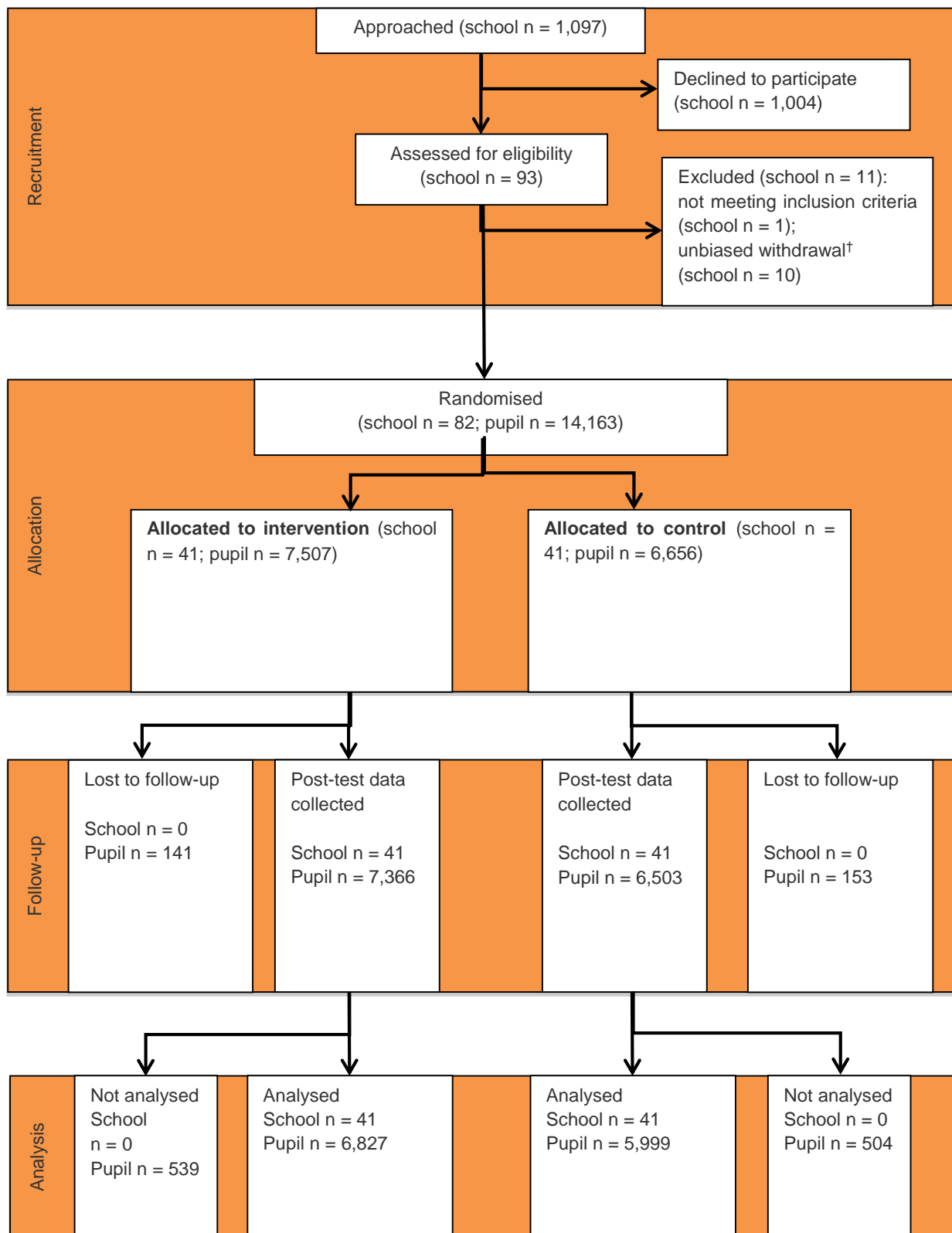
Just over 1,000 pupils from the 82 schools were lost to the analysis because they did not have a Key Stage 2 measure of prior attainment, likely because of being educated abroad or in the private sector at primary school. This was known at randomisation so can be regarded as unbiased attrition. Overall, the pupils in the analysis represented 90% of the pupils who were randomised.

The teacher-level experiment has a very different data structure, reflected in the participant flow diagram. The experiment also had more missing data because it relied on additional data and permission. The teacher-level experiment relied on being able to link pupils (and their outcomes) to their teacher using class lists provided by schools. All schools provided class lists for the first year of the study as it was a condition of randomisation, but only 48 schools provided them in the second year (24 intervention and 24 control).

We also required the schools’ permission to match pupils’ UPNs to their records in the NPD. This was sought in the MOU and therefore applied to all schools who remained involved in the evaluation throughout. However, when schools withdrew from the intervention or evaluation we sought the school’s permission to match the pupil data we had collected to the NPD. We did this because NFER regards withdrawal as withdrawal from the entire evaluation, including any data-matching permissions given,

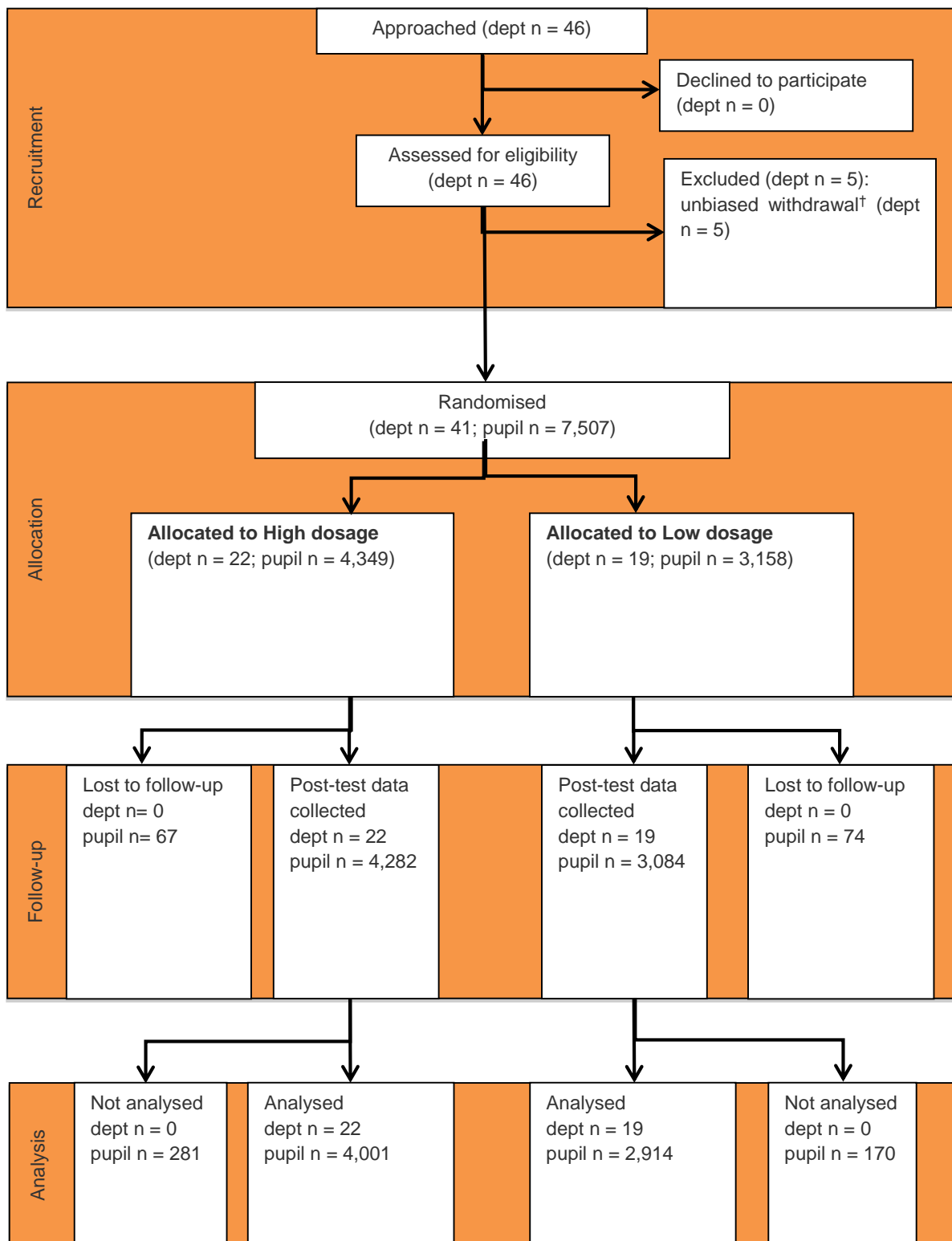
unless explicitly told otherwise. Of the 45 schools who withdrew over the course of the evaluation, 31 gave permission to match the pupil data, 13 refused, and one did not respond (we did not match that school's data). For these 14 schools, we used anonymized pupil outcome data from the NPD to analyse the school- and department-level experiments.

Figure 2: Participant flow diagram for the school-level experiment (Cohort 2, primary outcome)



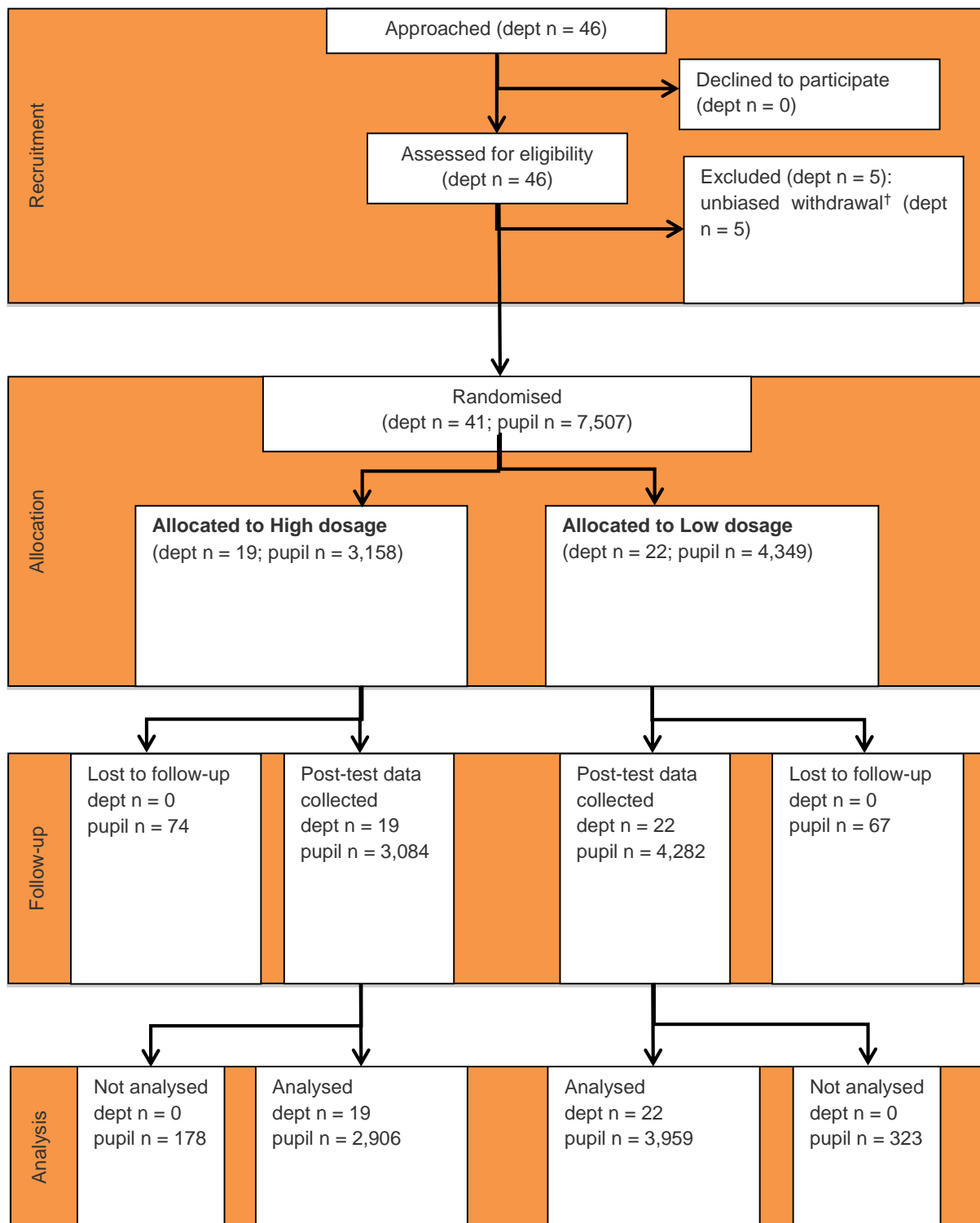
† Ten schools withdrew from participation in the evaluation before they were informed of their allocation to intervention or control. We retained the randomisation allocations for all schools and regarded this withdrawal from the evaluation as unbiased.

Figure 3: Participant flow diagram for the department-level experiment (maths departments)



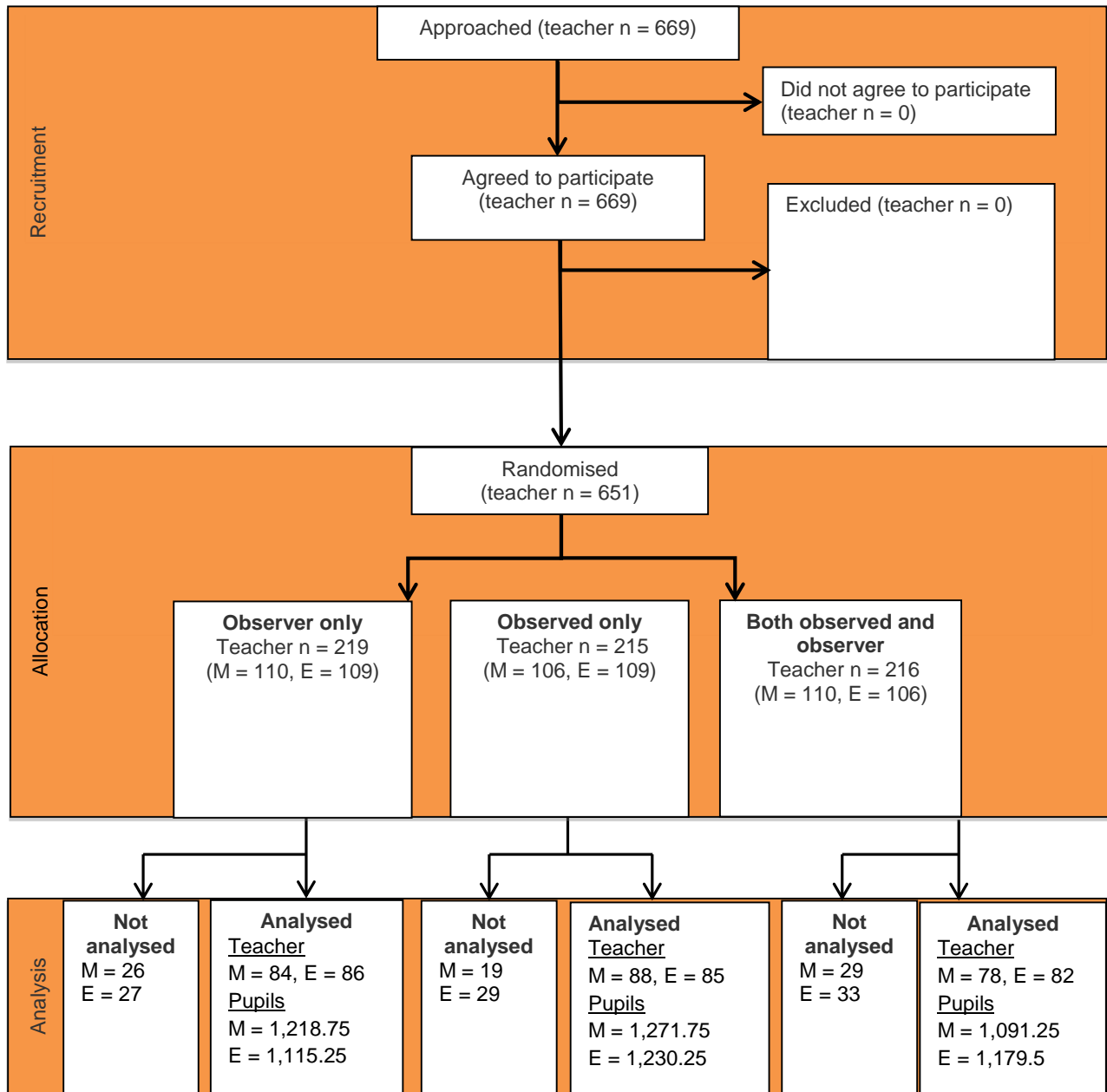
† Five intervention maths departments withdrew from participation in the evaluation before they were informed of their allocation to high or low dosage. We retained the randomisation allocations for all schools and regarded this withdrawal from the evaluation as unbiased. Four had been allocated to low dosage and one had been allocated to high dosage, which is why the number of departments analysed is unbalanced.

Figure 4: Participant flow diagram for the department-level experiment (English departments)



† Five intervention English departments withdrew from participation in the evaluation before they were informed of their allocation to high or low dosage. We retained the randomisation allocations for all schools and regarded this withdrawal from the evaluation as unbiased. Four had been allocated to high dosage and one had been allocated to low dosage, which is why the number of departments analysed is unbalanced.

Figure 5: Participant flow diagram for the teacher-level experiment



Note: E = English, M = Maths. Fractions of pupils are reported because pupils were taught by multiple teachers in the class lists provided by schools (shared classes) and many pupils had different teachers in the Year 10 class lists and the Year 11 class lists (changes of staffing). The sum of weights for each pupil in each subject across both years of the evaluation equalled one. However, the groups are determined by the pupils' teachers who may have been in different groups. Where a pupil has more than one teacher, a fraction of the pupil is allocated to each group that their teachers were allocated to.

Table 7 shows the minimum detectable effect size (MDES) for the school-level experiment at protocol stage, randomisation stage, and analysis stage. Although school recruitment missed the target, the loss of power was mitigated by a somewhat lower than expected ICC at analysis. The lower ICC might have been due to the specific eligibility criteria applied to schools at recruitment, making them more uniform.

Table 7: Minimum detectable effect size at different stages—school-level experiment

Stage	N [schools/pupils] (n = intervention; n = control)	Correlation between pre-test (+other covariates) & post-test	ICC	Blocking/ stratification or pair matching	Power	Alpha	Minimum detectable effect size (MDES)
Protocol	S = 120 (55; 65) P/S = 180	0.75	0.075	No stratification assumed	80%	0.05	0.10
Randomisation	S = 82 (41; 41) P/S = 173	0.75	0.075	No stratification assumed	80%	0.05	0.12
Analysis	S = 82 (41; 41) P/S = 156	0.67	0.046	No stratification assumed	80%	0.05	0.11

Note: the analysis included three stratification variables (see 'Randomisation' section) but, in order to be conservative, the sample size calculations at each stage did not include an assumption of the variance that might be explained by those stratification variables.

Table 8 shows the MDES for the department-level experiment at protocol stage, randomisation stage, and analysis stage. The department-level experiment always had less power than the school-level experiment, and missing the recruitment target, in combination with slightly higher than expected ICCs, exacerbated this.

Table 8: Minimum detectable effect size at different stages—department-level experiment

Stage	N [schools/pupils] (n=high dose; n=low dose)	Correlation between pre-test (+other covariates) & post-test	ICC	Blocking/ stratification or pair matching	Power	Alpha	Minimum detectable effect size (MDES)
Protocol	S = 55 (28; 27) P/S = 180	0.75	0.075	None	80%	0.05	0.15
Randomisation	S = 41 (22; 19) P/S = 183	0.75	0.075	None	80%	0.05	0.17
Analysis (maths)	S = 41 (22; 19) P/S = 169	0.68	0.087	None	80%	0.05	0.20
Analysis (English)	S = 41 (19; 22) P/S = 167	0.55	0.086	None	80%	0.05	0.23

Table 9 shows the MDES for the teacher-level experiment at protocol stage, randomisation stage, and analysis stage. The discrepancy between the ICC at protocol and analysis stages is due to the prevalence of setting. This was not taken into account in the original sample size calculations and is largely responsible for the increased MDES.

Table 9: Minimum detectable effect size at different stages—teacher-level experiment

Stage	N [teachers/pupils]	Correlation between pre-test (+other covariates) & post-test	ICC	Blocking/ stratification or pair matching	Power	Alpha	Minimum detectable effect size (MDES)
Protocol	T = 330 (110; 110; 110) P/T = 30	0.75	0.075	None	80%	0.05	0.08
Randomisation (maths)	T = 326 (110; 106; 110) P/T = 30	0.75	0.075	None	80%	0.05	0.08
Randomisation (English)	T = 324 (109; 109; 106) P/T = 30	0.75	0.075	None	80%	0.05	0.08
Analysis (maths)	T = 250 (84; 88; 78) P/T = 14	0.68	0.43	None	80%	0.05	0.19*
Analysis (English)	T = 253 (86; 85; 82) P/T = 14	0.58	0.29	None	80%	0.05	0.18*

*Calculated as per the analysis model i.e. comparison of two randomised groups versus one.

Pupil characteristics

Table 10 shows baseline characteristics of the 41 intervention schools and 41 control schools and their pupils. The difference in pre-test scores between intervention and control for the school-level experiment, expressed as an effect size from a multilevel model, is 0.01 (95% confidence interval: -0.08, 0.10).

Table 10: Baseline comparison for school-level experiment

Variable	Intervention group		Control group		
	School-level (categorical)	n/N (missing)	Percentage	n/N (missing)	Percentage
Academy		25/41 (0)	61%	26/41 (0)	63%
Ofsted rating:					
Outstanding		13/41 (0)	32%	6/41 (0)	15%
Good		15/41 (0)	37%	20/41 (0)	49%
Requires improvement		11/41 (0)	27%	13/41 (0)	32%
Inadequate		2/41 (0)	5%	2/41 (0)	5%

Performance stratifier (above median)	19/41 (0)	46%	19/41 (0)	46%
FSM stratifier (above median)	20/41 (0)	49%	19/41 (0)	46%
White ethnicity stratifier (above median)	20/41 (0)	49%	24/41 (0)	59%
School-level (continuous)	n (missing)	Mean	n (missing)	Mean
Number of Cohort 1 pupils (Y11 in 2014/15)	41 (0)	182	41 (0)	168
Number of Cohort 2 pupils (Y10 in 2014/15)	41 (0)	183	41 (0)	162
Number of Cohort 3 pupils (Y10 in 2015/16)	41 (0)	180	41 (0)	148
Pupil-level (categorical)	n/N (missing)	Percentage	n/N (missing)	Percentage
Ever eligible for FSM	2,767/6,827 (680)	41%	2,238/5,999 (657)	37%
Pupil-level (continuous)	n (missing)	Mean	n (missing)	Mean
Pre-test score	6,827 (680)	120.4	5,999 (657)	120.7

Table 11 shows baseline characteristics of the 19 and 22 schools in the department-level experiment. For the department-level experiment, the difference in pre-test scores is 0.02 (-0.09, 0.12).

Table 11: Baseline comparison for department-level experiment

Variable	High English/low maths		High maths/low English	
School-level (categorical)	n/N (missing)	Percentage	n/N (missing)	Percentage
Academy	12/19 (0)	63%	13/22 (0)	59%
Ofsted rating:				
Outstanding	7/19 (0)	37%	6/22 (0)	27%
Good	5/19 (0)	26%	10/22 (0)	45%
Requires improvement	6/19 (0)	32%	5/22 (0)	23%
Inadequate	1/19 (0)	5%	1/22 (0)	5%
School-level (continuous)	n (missing)	Mean	n (missing)	Mean
Number of Cohort 1 pupils (Y11 in 2014/2015)	19 (0)	192	22 (0)	169
Number of Cohort 2 pupils (Y10 in 2014/2015)	19 (0)	198	22 (0)	166
Number of Cohort 3 pupils (Y10 in 2015/2016)	19 (0)	194	22 (0)	164
Pupil-level (categorical)	n/N (missing)	Percentage	n/N (missing)	Percentage
Ever eligible for FSM	1,220/2,906 (178)	42%	1,569/3,959 (4)	40%
Pupil-level (continuous)	n (missing)	Mean	n (missing)	Mean

Pre-test score (English)	2,906 (178)	56.8	3,959 (323)	56.6
Pre-test score (maths)	2,914 (170)	63.6	4,001 (281)	63.0

Table 12 shows baseline pupil characteristics for the teacher-level experiment.

Table 12: Baseline comparison for teacher-level experiment

Variable	Observer only		Observed only		Both	
	n/N	pct	n/N	pct	n/N	pct
Pupil-level (categorical)						
Ever eligible for FSM	516/1,219	42%	504/1,272	39%	438/1,091	40%
Pupil-level (continuous)	n	Mean	n	Mean	n	Mean
Pre-test score (English)	1,219	64.6	1,271	64.1	1,091	65.1
Pre-test score (maths)	1,115	55.9	1,230	58.2	1,180	57.7

Outcomes and analysis

Table 13 shows the results from our analysis of the impact of the intervention on pupil outcomes. The effect size for the primary analysis of the impact of the intervention on 2016 GCSE English and maths is -0.01, which is equivalent to no difference between intervention and control in terms of months of progress. The difference is not statistically significant at the 5% level, suggesting the small between-group difference is likely to be due to chance. *The evidence from this analysis suggests that the Teacher Observation intervention had no impact on pupils' GCSE English and maths attainment compared to that of pupils in control schools.*

The analysis also suggests the intervention had no impact on 'ever FSM-eligible' pupils' GCSE English and maths attainment. The effect size for the FSM subgroup analysis is 0.01—also equivalent to no difference between intervention and control in terms of months of progress. The difference is not statistically significant. Analysis of the interaction between FSM eligibility and the intervention, measuring the extent of differential progress made by FSM-eligible pupils compared to non-FSM pupils, also showed no statistically significant difference.

Table 14 shows the results from our analysis of the impact of being allocated to high dosage (more observations) on pupil outcomes compared to intervention school departments that had a lower dosage (fewer observations). The 'Fidelity' section below shows that the dosage randomisation led to, on average, twice as many observation involvements per teacher in high-dose departments than in low-dose departments (12 vs 6 over two years). This suggests the dosage randomisation led to meaningful differences in the way the intervention was implemented. The effect size for the primary analysis of the dosage impact on GCSE English is -0.03 and for GCSE maths is 0.07. These effects are equivalent to one less month of progress made by pupils in high dosage English departments compared to low dosage ones, and one additional month of progress in high dosage maths departments compared to low dosage ones. However, neither difference is statistically significant. From this analysis there is no evidence of an impact on pupils' GCSE English and maths attainment in departments selected to receive a higher dosage of Teacher Observation compared to low-dosage departments. There was also no evidence of an impact of dosage on FSM-eligible pupils' GCSE English and maths attainment.

Table 15 shows the results from our analysis of the impact of teachers being the one observing compared to not observing, and being observed compared to not being observed, on the attainment of pupils they taught. None of the differences are statistically significant. There is no evidence of a differential impact on pupils' GCSE English and maths attainment depending on whether a teacher

observes or is observed. Including data from the control group (counting control teachers as a group that did no observing and were not observed) gave very similar results to estimates based only on differences between types of intervention teacher.

Table 16 summarises the results from an on-treatment analysis of the impact of the number of observations each teacher completed and the attainment of pupils they taught. Since the number of observations completed was in the control of the individual teacher, this analysis cannot be regarded as causal. However, differences may be indicative of greater impact from more engagement with peer observation. The effect sizes are presented as the association between attainment and the number of observations, expressed as the expected effect on attainment of increasing the number of observations from the lower quartile (25th percentile of teachers) to the upper quartile (75th percentile of teachers). Neither of the associations between the overall number of observations completed and attainment are statistically significant. Splitting the on-treatment analysis into, respectively, the number of observations as the observer and the number of observations as the one observed also found no positive associations with attainment (the association between being observed and pupils' maths GCSE attainment was negative and significant at the 5% level). There is no evidence from this analysis of a positive association between the number of observations a teacher completed and the GCSE English and maths attainment of the pupils they taught.

A number of secondary analyses were undertaken in addition to the primary analyses for each experiment (results shown in Tables 13–15) which measured the impact on different cohorts, individual subjects, and on Year 10 test outcomes. Across these 24 analyses summarised in Tables 13–15 there was one positive significant effect at the 5% level (intervention impact on 2015 GCSE English) and 23 differences that were not statistically significant. However, as no consistent pattern in effect sizes was evident across the analyses, the positive finding is not compelling evidence that the intervention had an impact in that particular case. Indeed, under the assumption that there is no underlying 'true' effect, there is a 71% probability of one or more of the 24 analyses showing a significant effect at the 5% level.

Table 13: School-level experiment impact analysis

Outcome	Raw means				Effect size		
	Intervention group		Control group		n in model (intervention; control)	Hedges g (95% CI)	p- value
n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)				
Primary Outcome: Cohort 2, 2016 GCSE English + maths	7366 (141)	76.7 (76.2, 77.2)	6503 (153)	76.1 (75.6, 76.6)	12826 (6827; 5999)	-0.01 (-0.08, 0.06)	0.802
FSM-only subgroup: Cohort 2, 2016 GCSE English + maths	2992 (104)	70.1 (69.3, 71.0)	2458 (103)	68.8 (67.9, 69.6)	5005 (2767, 2238)	0.01 (-0.08, 0.10)	0.77
Cohort 1: 2015 GCSE English	7363 (88)	39.0 (38.7, 39.2)	6843 (59)	38.0 (37.7, 38.2)	14147 (7327; 6820)	0.11 (0.01, 0.21)	0.03
Cohort 2: 2016 GCSE English	7366 (141)	38.9 (38.6, 39.1)	6503 (153)	38.5 (38.2, 38.7)	12920 (6865, 6055)	0.01 (-0.07, 0.09)	0.819
Cohort 1: 2015 GCSE maths	7362 (88)	37.4 (37.1, 37.7)	6843 (59)	36.9 (36.6, 37.2)	14147 (7327, 6820)	0.06 (-0.02, 0.14)	0.128
Cohort 2: 2016 GCSE maths	7366 (141)	37.8 (37.5, 38.1)	6503 (153)	37.6 (37.3, 37.9)	12990 (6915, 6075)	-0.02 (-0.08, 0.05)	0.623
Cohort 2: 2015 Year 10 reading test	1643 (2111)	13.5 (13.2, 13.9)	1602 (1726)	14.7 (14.3, 15.0)	3067 (1546, 1521)	-0.15 (-0.31, 0.02)	0.086
Cohort 3: 2016 Year 10 reading test	1498 (2087)	13.5 (13.2, 13.9)	975 (2049)	13.1 (12.6, 13.5)	2313 (1408, 905)	0.03 (-0.18, 0.23)	0.814
Cohort 2: 2015 Year 10 maths test	1849 (1905)	24.7 (24.0, 25.4)	1676 (1652)	25.8 (25.0, 26.6)	3354 (1758, 1596)	-0.09 (-0.22, 0.03)	0.136
Cohort 3: 2016 Year 10 maths test	1496 (2193)	25.1 (24.3, 25.9)	1016 (2008)	24.4 (23.4, 25.4)	2375 (1408, 967)	-0.05 (-0.17, 0.07)	0.448

Table 14: Department-level experiment impact analysis

Outcome	Raw means				Effect size		
	High dosage group		Low dosage group		n in model (intervention; control)	Hedges g (95% CI)	p- value
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)			
Primary Outcome: Cohort 2, 2016 GCSE English	3084 (74)	38.8 (38.5, 39.2)	4282 (67)	38.9 (38.5, 39.1)	6865 (2906, 3959)	-0.03 (-0.18, 0.12)	0.713
Primary Outcome: Cohort 2, 2016 GCSE maths	4282 (67)	38.1 (37.8, 38.5)	3084 (74)	37.4 (36.9, 37.8)	6915 (4001, 2914)	0.07 (-0.06, 0.20)	0.288
FSM-only subgroup: Cohort 2, 2016 GCSE English	1304 (59)	35.9 (35.3, 36.5)	1688 (45)	35.9 (35.4, 36.5)	2789 (1220, 1569)	0.02 (-0.14, 0.17)	0.836
FSM-only subgroup: Cohort 2, 2016 GCSE maths	1688 (45)	34.6 (34.0, 35.3)	1304 (59)	33.6 (32.8, 34.4)	2810 (1587, 1223)	0.05 (-0.10, 0.20)	0.506
Cohort 1: 2015 GCSE English	3172 (46)	38.7 (38.3, 39.1)	4190 (42)	39.2 (38.9, 39.5)	7327 (3160, 4167)	-0.06 (-0.20, 0.08)	0.400
Cohort 1: 2015 GCSE maths	4190 (42)	37.3 (36.9, 37.7)	3172 (46)	37.5 (37.0, 37.9)	7327 (4167, 3160)	-0.01 (-0.15, 0.12)	0.984
Cohort 2: 2015 Year 10 reading test	666 (913)	13.5 (12.9, 14.1)	977 (1197)	13.6 (13.1, 14.0)	1546 (626, 920)	-0.08 (-0.39, 0.24)	0.643
Cohort 3: 2016 Year 10 reading test	531 (1028)	12.8 (12.2, 13.4)	967 (1162)	14.0 (13.5, 14.4)	1408 (492, 916)	-0.27 (-0.63, 0.08)	0.144
Cohort 2: 2015 Year 10 maths test	1080 (1094)	24.0 (23.1, 24.9)	769 (810)	25.7 (24.6, 26.8)	1758 (1020, 738)	-0.03 (-0.26, 0.21)	0.812
Cohort 3: 2016 Year 10 maths test	989 (1140)	24.6 (23.6, 25.6)	507 (1052)	25.9 (24.6, 27.3)	1408 (937, 471)	-0.03 (-0.27, 0.22)	0.837

Table 15: Teacher-level experiment impact analysis

Outcome	Effect size		
	Weighted n in model (observer; observed; both)	Hedges g (95% CI)	p-value
Observer vs non-observer: 2016 maths GCSE	3820.25 (1218.75, 1271.75, 1091.5)	0.10 (-0.49, 0.68)	0.747
Observed vs non-observed: 2016 maths GCSE	3820.25 (1218.75, 1271.75, 1091.5)	0.01 (-0.58, 0.61)	0.963
Observer vs non-observer: 2016 English GCSE	3793 (1115.25, 1230.25, 1179.5)	-0.02 (-0.50, 0.47)	0.948
Observed vs non-observed: 2016 English GCSE	3793 (1115.25, 1230.25, 1179.5)	0.14 (-0.35, 0.62)	0.586

Table 16: Teacher-level on-treatment analysis

Outcome	Effect size		
	Interquartile range: number of observations (25 th , 75 th percentile)	Hedges g (95% CI)	p-value
Association between 19 additional observations and pupils' maths GCSE attainment	19 (1, 20)	-0.06 (-0.13, 0.02)	0.144
Association between 25 additional observations and pupils' English GCSE attainment	25 (0, 25)	0.04 (-0.04, 0.13)	0.333
Association between 12 additional observations <i>being observed</i> and pupils' maths GCSE attainment	12 (0, 12)	-0.10 (-0.19, -0.02)	0.022
Association between 10 additional observations <i>as the observer</i> and pupils' maths GCSE attainment	10 (0, 10)	0.02 (-0.05, 0.08)	0.630
Association between 9 additional observations <i>being observed</i> and pupils' English GCSE attainment	9 (0, 9)	0.02 (-0.06, 0.11)	0.622
Association between 8 additional observations <i>as the observer</i> and pupils' English GCSE attainment	8 (0, 8)	0.02 (-0.05, 0.10)	0.539

Cost

The Teacher Observation intervention had many bespoke elements which were put together in the unique context of implementing the intervention for an EEF project. The cost of the intervention for this evaluation, presented below, has been estimated on the basis of costs that were part of the project. While it provides a rough estimate of the cost that might be expected, the costs do not necessarily represent what the cost would be to a school if it were to deliver this intervention independently.

This cost evaluation estimates the cost to schools under the assumption that the EEF funding for financial costs, prerequisite costs, and compensation for staff time is not being provided and that schools are paying for their share of the total costs they would otherwise bear.

Financial costs

The main financial costs of delivering the intervention, which were borne by EEF as part of this project, were the RANDA TOWER software licences and the costs associated with delivering training.

RANDA TOWER software was used in the intervention to support staff carrying out observations. A licence to use the RANDA software across all schools in the intervention was negotiated with RANDA Solutions, an education technology company based in the United States. The cost of £200,000 for this set of licences covered two years of use for the 41 intervention schools. This translates to an ongoing cost of £2,500 per school per year. This cost was unique to the project and the cost for a school or group of schools obtaining software licences would need to be negotiated directly with the supplier. The software also incorporated a bespoke rubric that was developed especially for the project.

Table 17 presents an analysis of the one-off budgeted costs associated with training divided by the number of intervention schools that were assumed in the budget. The range of staff costs presented is due to uncertainty as to whether particular staff costs relate to delivering the training (within the scope of this cost evaluation) or developing the intervention (outside the scope of this cost evaluation). The cost of training is a one-off cost incurred at the beginning of the evaluation. The training scheme was prepared and delivered by CMPO especially for this project and is not commercially available.

Table 17: CMPO's financial costs per school for training

Cost item	Cost per school
Visits to schools	£800
Printing of materials for schools, postage, and other office costs	£87
Travel	£127
Staffing	Midpoint cost: £2,453 (Range: £1,239–£3,668)
Total	£3,468 (Range £2,253–£4,683)

As part of the process evaluation we asked schools whether they had incurred additional costs in implementing the intervention. We received few responses to these survey questions which could indicate that there was little additional financial expenditure required for schools. Those that did respond cited the costs of refreshments and printing for training and feedback sessions amounting to £300 per school. We include this in our estimate of total cost, although it is a conservative estimate given the limited information it is based on.

We combined the one-off financial costs (training) with the ongoing costs (software licence and other costs) to estimate the total cost per school over three years to reflect the cost of implementing the intervention over a sustained period. Table 18 shows our estimate of the cost per school over time, the cumulative cost per school over three years, and the average cost per school per year. The average cost per school per year was roughly £4,000.

Table 18: Average financial cost per school per year

Year	Cost per school	Cumulative cost per school	Average cost per school per year
First year	£6,177	£6,177	£6,177
Second year	£2,709	£8,887	£4,443
Third year	£2,709	£11,596	£3,865

We derived the cost per pupil by dividing the average cost per school per year by the average number of pupils per school. The evaluation focused on the attainment outcomes of three cohorts of Year 10 and Year 11 pupils, but as the intervention is a whole-school intervention, the number of pupils that would be influenced by the intervention activities is likely to be all pupils in a school over a three-year period. Based on a typical secondary school intake of 180 pupils and a total of seven cohorts of pupils over three years (Year 7–11 pupils, plus two new cohorts of Year 7s), we divide the cost by 1,260 pupils. This calculation yields a total cost per pupil over three years of £8.28 and an average cost per pupil per year of £2.76.

Regardless of the assumptions made on the cost of training, this intervention has a very low financial cost per pupil.

Staff time

The Teacher Observation programme involved a day's training in using the software and the rubric for one member of staff in the school. The lead staff member then trained and supported other staff members in implementing the intervention. Schools implementing the intervention therefore needed to allocate several days of training time for the lead staff member and additional staff time for the information to be shared. Some supply cover is likely to have been necessary to accommodate this training.

The amount of staff time required to complete the observations would depend on the level of implementation. On average, teachers completed around ten observations each over two years, while 25 observations per teacher were completed in some intervention schools. The amount of supply cover necessary to allow staff time to complete observations would depend on timetabling, that is, whether teachers are able to observe other teachers during their free periods. Many schools found that regular supply cover was necessary to enable observations because observations were within departments, and departments had timetables that meant staff had lessons and free periods at the same time.

Prerequisite costs

The RANDA TOWER software used in the intervention required an observer to have an iPad and the rubric was loaded on to this as an app. iPads were provided by CMPO and purchased by the EEF for this evaluation. Schools would require the technology necessary for using the software in this way to implement the intervention; this may require additional investment for a school with limited IT resources.

Process evaluation

Summary of Process evaluation findings

- Maths and English departments in each school were allocated to either high or low dosage observations. The minimum number of observations was set at three in the low dosage departments and four in high dosage departments.
- Teachers from both the control and intervention groups reported that their students were used to having observers in class.
- In-school training was cascaded by project co-ordinators and heads of departments via formal or informal meetings or through written information. The majority of *teachers reported that they felt sufficiently prepared* for the intervention, and were content with the level of training they received.
- Most teachers in English and maths departments in intervention schools were involved, to some extent, in observations across the two-year period. However, the final dosages were around a quarter of the developers' initial expectations; 60% of respondents felt that the project was more difficult to organise than expected and 50% felt that their *current level of observation was not sustainable*.
- Around half the observers were happy to use their free or non-contact time to conduct the observations, and many preferred to do this than take time out from their teaching.
- Over half the teachers *felt uncomfortable taking time out of teaching to complete observations*.
- Teacher engagement within the intervention group varied greatly. This appeared to relate to the enthusiasm or support offered by the project co-ordinator or the extent to which the intervention was viewed as valuable CPD. Around a third of project co-ordinators felt that the project schedule was more demanding than expected and that it was difficult to engage staff.
- Observations were more likely to be completed where they had been actively timetabled by the school co-ordinator.
- The project appeared to run most successfully when participants viewed it as peer-driven CPD and formal planning and reflection sessions were scheduled to complement the observation sessions.
- Almost half the teachers reported *some difficulties with the technology*.
- The majority of observers agreed that the *observation software and tablet were user-friendly and intuitive*.
- In terms of fidelity, teachers broadly adhered to their allocated high or low dosage; high-dosage departments completed around twice as many observations, on average, as low-dosage departments. However, the total mean number of observations was below the developers' initial expectations and resulted in an average of three observations per observer in low dosage departments and six in high dosage departments over two years, as opposed to six and 12 observations per annum for observing teachers as originally recommended.
- Teachers were generally evenly divided between positive and negative views in terms of their general attitudes, their engagement with the programme, and their opinions about the perceived outcomes and benefits of the programme.

With regard to formative findings:

- It seems likely that teacher ‘buy-in’ would play a significant role in the successful implementation of an intervention of this kind.
- The co-ordinator appears to have an important impact on teachers’ attitudes and willingness to become (and stay) involved in the Teacher Observation programme.
- The way in which the intervention is introduced to colleagues seems to be an important factor—sharing of information, discussing the purpose of the observations, and what participants hope to gain from them.
- The level of support and encouragement the co-ordinator offers, particularly in terms of timetabling and cover, was also regarded by teachers as an important enabling factor

Control group activity:

- Control group schools revealed some existing peer observation activities and did not report any changes to their approaches to teaching maths or English in the last year that might affect the results of the impact evaluation.
- It is possible that the relatively small proportion of ‘additional’ observations in the intervention group were not enough to make an impact. Indeed, some intervention schools reported that they had simply adapted their ‘business as usual’ peer observations to incorporate the use of the RANDA software.

Process evaluation discussion

The process evaluation for the Teacher Observation project focused on understanding how teachers and schools implemented and engaged with the intervention, whether the required observation schedules could be practically maintained, and how they were used within schools. We also explored perceived impacts in terms of any changes in teachers’ practices or impact on the students’ learning.

The process evaluation included the following elements:

- observation of the CMPO-led training on the use of the RANDA TOWER software;
- a teacher survey of participating teachers (intervention and control) and project co-ordinators;
- case study visits to six schools, selected to give a range of school type, geographical spread, and covering high and low dosage in English and maths departments;
- analysis of RANDA TOWER software;
- follow-up telephone interviews to six school co-ordinators (intervention schools); and
- follow-up focused pro forma open questions to all intervention schools.

The findings are summarised in this section. We present key findings from the teacher survey and, where relevant, illustrate them with examples from the case study visits to intervention schools (in text boxes) or themes or quotations from the interviews and pro formas.

In June 2015, we invited maths and English teachers in all participating schools to complete the survey, with additional questions for heads of maths or English Departments, or project co-ordinators. The survey for the control group was considerably shorter, although there were some overlapping questions, as discussed below.

We received 264 completed survey questionnaires (Table 19). The proportions in both intervention and control groups were broadly 55–60% from maths departments and 40–45% from English departments.

Table 19: Survey respondents by group

	Control group (N)	Intervention group (N)
Year 10/11 teachers	119	83
Project co-ordinator/ head of department	39 (83%)	23 (56%)
Total	158	106

The majority of project co-ordinators and heads of departments (HoDs) across both groups reported that before the Teacher Observation project, peer observation had been used in their school (Table 20). This suggests that the intervention may have been a formalisation of existing practice—using the RANDA software to record observations that would have been done anyway rather than distinct practice. However, the intervention may have encouraged teachers to conduct more observations than they would have done otherwise.

Table 20: Previous use of peer observation in participating schools

		Control group (N)	Intervention group (N)
Prior to the Teacher Observation project, was peer observation used in your school?	Yes	27	22
	No	11	1
	missing	1	0
Total		39	23

The data in Table 21 suggests that having an observer in class is not unusual across secondary schools generally. Project co-ordinators in both intervention and control groups reported that where peer observation had taken place, it had most commonly been used for performance management purposes but had also been used quite frequently for CPD purposes. Teachers were more likely to have been observed by someone within their department than from a different department, although some cross-departmental observations were recorded. ‘Open lessons for purposes of continuing professional development (CPD)’ were less frequently reported, as were situations where only senior management carried out the observations.

Teachers from both the control and intervention groups also reported that their students were used to having observers in class (see Table 22). The majority of respondents in both groups felt that observing others was more useful than being observed. As might be expected, a higher proportion of respondents in the intervention group—when compared to the control group—reported that being involved in the project had made them reflect on their practice. Finally, both intervention and control group respondents reported that they would have preferred more opportunities to observe colleagues.

Table 21: Previous use of peer observation in participating schools (project co-ordinators)

	Most teachers have been observed (N)		Some teachers have been observed (N)		Most teachers have conducted observations (N)		Some teachers have conducted observations (N)		Only senior management have conducted observations (N)		N/A (N)	
	Control	Intervention group	Control	Intervention group	Control	Intervention group	Control	Intervention group	Control	Intervention group	Control	Intervention group
Peer observations <i>within</i> departments for CPD purposes	14	14	10	5	6	3	10	5	3	0	1	0
Peer observations <i>across</i> departments for CPD purposes	4	4	16	10	4	1	12	10	3	2	1	1
Open lessons for CPD purposes*	1	3	9	10	3	2	8	8	2	0	7	1
Observations for performance management purposes	25	13	0	5	1	0	10	8	1	3	0	0

Control (N = 27) and intervention group (N = 22).

* For example, when teachers are invited to visit specific lessons delivered by colleagues.

Table 22: Teachers' attitudes to peer observation

	Strongly disagree (%)		Disagree (%)		Agree (%)		Strongly agree (%)		Don't know (%)	
	Control	Intervention group	Control	Intervention group	Control	Intervention group	Control	Intervention group	Control	Intervention group
I think being observed is more useful than observing others.	10	10	63	67	15	7	4	2	7	15
My students are used to having observers in class.	0	0	11	2	63	62	24	34	3	2
Being involved in this project has made me reflect on my own practice.	19	8	28	24	26	60	3	7	24	2
I would have preferred (more) opportunities to observe colleagues.	3	4	26	37	46	46	17	10	9	1

Control n ranged from 155–156 and intervention group n ranged from 67–105.

Implementation

At the start of the intervention, two teachers (typically the project co-ordinators or heads of English or maths departments) received training from CMPO. For other teachers, training was cascaded by project co-ordinators via formal or informal meetings or through written information. Each observer had individual login access to iPads with the RANDA TOWER software installed. Teachers in participating schools were expected to complete the appropriate minimum number of observations according to their department's allocated dosage.

Preparation

The majority of teachers reported that they felt sufficiently prepared for the intervention, with three-quarters of teachers feeling content with the level of training they received.

Eighty-one per cent of teachers said that they felt adequately briefed about what they had to do in the project and 63% felt adequately briefed about the purpose of the intervention. The majority of observers agreed that the observation software and tablet were user-friendly and intuitive (89%). These findings were confirmed during case study interviews with almost all teachers reporting that the software was intuitive and easy to use.

Eighty per cent of respondents said that they were able to conduct their part of the project as intended; 88% of respondents also said that they were happy to take part in the project.

However, a number of teachers also reported some concerns:

- 60% of respondents felt that the project was more difficult to organise than they expected;
- 50% of intervention respondents felt that their current level of observation was not sustainable (and a further 10% did not know); and
- 16% of respondents said that they felt that insufficient training had been a barrier to implementation.

The feasibility study conducted by NFER at the outset of the evaluation identified the difficulty of organising the observations (particularly the complexity of timetabling and organising cover), as well as insufficient teacher training, as risks for implementation. The findings from the feasibility study are shown in Appendix D.

Case study interviews suggested that, where time had been invested in informing staff about the purpose of the project and involving them in discussion and planning, they were generally enthusiastic and engaged.

'[The co-ordinator] was pretty good at explaining the purpose of things at the beginning. Our meetings were front loaded, and we had quite a few before we physically started doing the observations; that was good. We talked about the descriptors, and clarified our understanding. We also discussed how we would use the observations to support our ongoing CPD and that was really useful (English teacher, School 3).

In some case study schools, however, some members of staff were openly disengaged, reporting that they knew nothing about the purpose of the project or about why they were being asked to take part in observations. In these schools, staff felt they had been imposed upon and were disinclined to ensure that their allocated observations were completed.

'There was no prior consultation; we were just told we were doing it so most of us are just "going through the motions" to complete the observations. Nobody is doing anything with the results. We have no time to discuss anything. The project rather fell flat—no one was very interested—we have several other things to do and I would

rather spend time teaching my class! We weren't really given enough information—I would have liked to know the research rationale. So there was no buy-in and no additional time set aside to discuss anything' (maths teacher, School 2).

The way in which the intervention is introduced and supported and whether the observations are used to support professional development in the school seem to be crucial to teacher engagement and, therefore, conditions for success.

Cover

In terms of covering teachers' time to conduct the required observations, almost 30% of respondents completed observations during their non-contact time only. For 14% of respondents, cover was arranged through normal in-school cover arrangements (for example teaching/curriculum assistants). Almost half of respondents (49%) used a mixture of these approaches, with a minority of 6% of respondents reporting that supply cover was specifically arranged for the project. This pattern of cover was also reflected in the case study visits.

Fifty-three per cent of teacher observers said they were happy to use their free or non-contact time to conduct the observations, but 38% disagreed or strongly disagreed with this statement.

Year 2 pro forma returns indicated that in most cases, subject departments were responsible for arranging cover, with senior managers only stepping in when needed.

Case study visits revealed differences in levels of support offered by the in-school co-ordinator which seemed to relate to teacher engagement. One teacher encapsulated a common theme when she said:

'Having a good co-ordinator is key. The co-ordinator made sure that all teachers felt comfortable with their observation partners. She reminded everyone when the observations were due to take place and arranged cover via the "curriculum assistants programme" in school. She made it easy for us to do what was needed' (maths teacher, School 1).

In this case, teaching assistants (TAs) were assigned to classes, as required, to cover for parts of lessons while the class teacher went out to do their observations. In this way, the class teacher was able to take the majority of his or her class as usual, but only lost a 20-minute slot when they were completing their observation. Teachers in this school appeared fully engaged and enthusiastic about the intervention.

In a different school, however, one teacher commented that:

'To be honest, we have just been told we have to do it. I feel we have other more important things to spend our time on. We already do planned observations, and it is really time consuming to have to arrange to do them yourself' (English teacher, School 5).

The evidence across the two years was mixed as to whether teachers from maths departments were more or less engaged with the initiative than their colleagues in English. For example, sub-group analyses of the Year 1 survey data suggested that English teachers were generally more positive about the initiative than maths teachers, suggesting they were also more engaged. By contrast, Year 2 returns suggested the opposite, with most respondents indicating that 'some' teachers from participating English departments were 'fully engaged', while 'most' maths teachers were 'fully engaged'. Some caution should be taken in interpreting these responses due to the small sample sizes involved.

Timetabling

While 51% of respondents agreed with the statement, ‘the timetabling of observations was done by our project co-ordinator’, 44% disagreed or strongly disagreed with this. Around half of respondents (49%) did, though, say that arranging cover for observation sessions was straightforward.

Evidence from the case study visits indicated that teachers were happier, and more inclined to complete their observations when a project co-ordinator set up the timetable and arranged cover:

‘It was made easy for me to do my observations. [The project co-ordinator] set up the pairings and arranged the timetable—I just had to be in the right place as instructed. It would have been much harder if she hadn’t been so diligent and encouraging’ (English teacher, School 1).

In two schools we visited, teachers were expected to make their own arrangements to observe colleagues and to arrange cover for observations. Teachers in those schools reported feeling burdened in a way that other teachers (whose observation slots had been allocated by the co-ordinator) did not:

‘We are expected to liaise with our colleagues about when we will be observing, and quite frankly, I have so many other priorities this tends to take a bit of a back seat. Then it turns into a pressure—I could do without it really’ (maths teacher, School 2).

Year 2 pro forma returns further confirmed that schools’ approaches to timetabling the observations were varied, with some schools devolving this to subject departments, and others using the project co-ordinator to lead this.

Completing observations

Of teacher observers, 71% felt that fitting in the required number of observations was difficult. In addition, teachers were not confident that completing the observations was a good use of their time. Over half the teachers (55%) felt uncomfortable taking time out of teaching to complete observations and 19% of teachers felt that other aspects of their work had suffered because of their involvement in the project.

Although many teachers spoke enthusiastically about being involved in the peer observation project, many, especially in high dosage departments, also expressed concerns that taking time out from teaching their GCSE students to complete the observations could be detrimental to their learning in this crucial year. Some suggested that the observations would be more beneficial (to them) if they could be conducted during the Year 9 maths/English classes:

‘We are only doing them when we are not teaching; we haven’t had ourselves covered for them because we just can’t justify that, we can’t miss 20 minutes of our own lesson to go and observe someone else, unfortunately, so we have been doing them in our free periods’ (maths teacher, School 6).

Barriers

Timetabling issues and staffing were most commonly mentioned as barriers to conducting the project as planned, cited by 68% of project co-ordinators and 58% of heads of departments. Changes in staffing and timetable clashes were also mentioned in the case studies.

Almost half (47%) of respondents reported difficulties with the technology and 42% reported that problems with software had been a barrier. It seems likely that these issues were technical in nature as 89% of observers agreed that the observation software and tablet were user-friendly and intuitive. For

some, the technology still presented challenges: the tablet application repeatedly failed to synchronise, and the use of passwords and IDs proved problematic for staff.

Case study interviews suggested that a number of participants had trouble logging on for the first time, but these respondents reported that, once they had begun using the software to conduct observations, it was very straightforward. Most co-ordinators, however, told us that they had suffered glitches when uploading the results, sometimes losing records completely necessitating repeat observations which had not been scheduled. One co-ordinator said:

'Synchronisation was a major issue. It may be a Wi-Fi issue, but I wasted such a lot of time on that' (school co-ordinator, School 5).

As with the case study findings, Year 2 pro forma returns indicated that the main drawbacks of the initiative related to time and technology. In terms of time, several schools found it challenging trying to schedule observations into their busy curriculum programmes. In addition, timetabling restrictions limited who could observe whom. At least one school compensated study participants by reducing the number of hours teachers had to participate in other CPD activities. Around a third of project co-ordinators felt that the project schedule was more demanding than expected and that it was difficult to engage staff (37% in both English and maths departments) and cited these as barriers to implementation.

Other reported drawbacks reported in Year 2 included staffing changes during the research period impacting on the consistency of the intervention, and the targeting of exam classes taking class teachers away from students at a key time in the academic year.

Conditions for success

As well as having an active co-ordinator arranging timetabling and cover, the ways in which the observations were used within the school was also found to relate to teacher engagement.

The extent of pre-planning and debriefing in relation to observations varied considerably. In some schools, formal planning and reflection sessions were scheduled as part of whole-school CPD:

'We weaved the project into our CPD programme and coaching pathways. We found that it supported good discussions about teaching and learning. Teachers at the beginning of their careers, especially, learnt/observed additional behaviour for learning strategies' (school co-ordinator, School 4).

'We agreed that the feedback would be narrative and discursive and sensitive pairings meant that it was always received as constructive and collaborative. Most teachers had informal discussions before and after observations and it was very much viewed as peer driven cpd. We will have a slot in our end of year departmental meetings where we discuss what we have learned and identify ways we might be able to use the software more effectively' (English teacher, School 1).

In a number of schools, feedback was available if requested; in others, teachers reported informal discussions between colleagues:

'I would like more time to feedback to anyone observed. I therefore think in one year, eight is acceptable (one per half term) in addition to the ordinary observations that take place. You need at least two hours per observation to do it thoroughly' (English teacher School 6).

'Feedback and dialogue would be important with this kind of programme – but we have not had time to do that' (maths teacher, School 2).

In one school, disengaged teachers reported just ‘going through the motions’, and in another, teachers confirmed the project co-ordinators statement that:

‘We didn’t do that much prep in school; we discussed how to use the software, but didn’t formally discuss (as a department or staff) how to make the ratings or use the observations—mainly because of time constraints, but also lack of interest. Similarly, [there was] no discussion following the observations as a department or staff team’ (school co-ordinator, School 5).

The project appeared to run most successfully when participants viewed it as peer-driven CPD.

Fidelity

In addition to the implementation issues explored through the questionnaire and case study interviews, we analysed the RANDA TOWER software data provided by CMPO to explore the fidelity to the intervention.

In Year 1, the intervention was slow to get underway. Early viewings of the data at the end of November 2014 showed that only one observation had been completed in one school. CMPO explained that many schools did not receive their iPads and login details until the second half of the autumn term and therefore could not have started until November. By March, CMPO were concerned at the rate of completion of the observations as only 351 had been completed—21% of the expected target at that point in the project. Teachers reported to CMPO various reasons for the delays. These included:

- getting over the ‘first observation hurdle’—teachers were apprehensive about starting the observations and how difficult they would be, but once they started these concerns were alleviated;
- staffing issues—this included key staff leaving or being away or ill;
- technical issues—teachers reported having trouble syncing the iPads or not clicking ‘complete’ meaning that the observations were still showing as ‘in progress’ (in response, CMPO sent some additional guidance about these issues to schools in March and helped to solve problems when they occurred); and
- confusion about who had responsibility for co-ordinating the observations within the school, and communication difficulties between observers and observees; in response, CMPO provided further advice to the project co-ordinators on how to organise the observations and to remind them of their responsibility to help keep the project running smoothly (in some instances, CMPO set up a draft observation schedule for schools that had not got one in March/April 2015).

Figure 6 shows how the number of observations being completed in intervention schools evolved over time. By April 2015, the rate of observations had increased and 760 observations had been completed. This was a significant increase and the rate of observations continued to rise steadily until the end of the academic year. At the end of the first year of the intervention, the RANDA TOWER data contained 1,873 records from 34 intervention schools, each of which included the unique ID of the teacher that was observing (observer) and the unique ID of the teacher that was observed (observee). By the end of the second year of the intervention, the RANDA TOWER data contained 6,276 records from 34 intervention schools: 3,138 observations in total.

Figure 6: Total number of observations completed over time

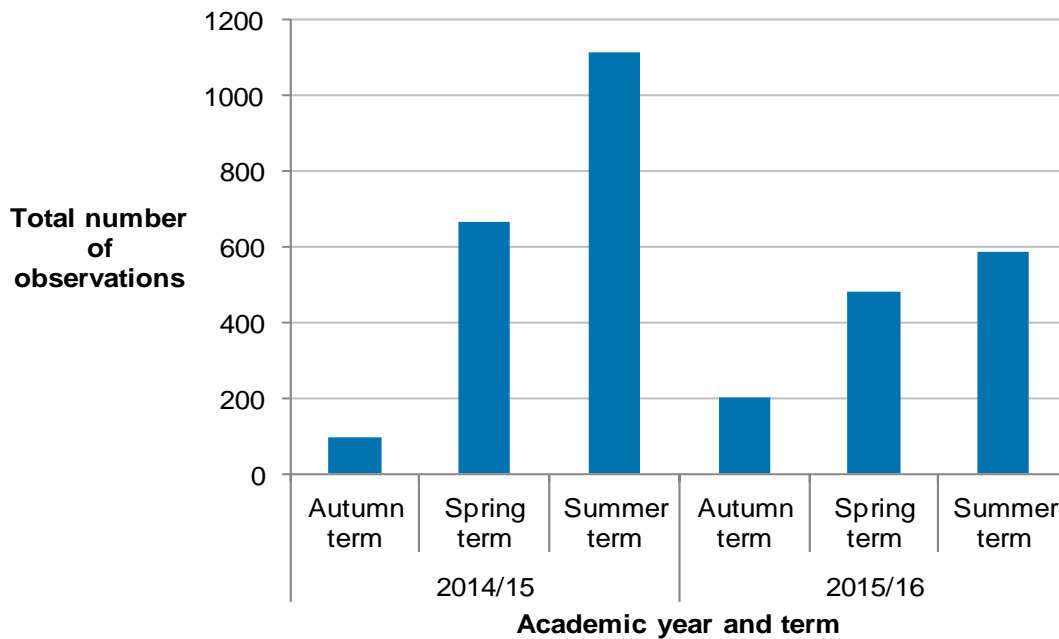


Figure 7 shows the proportion of teachers who completed at least one observation in each school. Seven out of the 41 intervention schools did no observations at all, whereas all teachers who were intended to participate were involved in observations in nine out of 41 intervention schools.

Figure 7: Proportion of teachers who completed at least one observation, by school

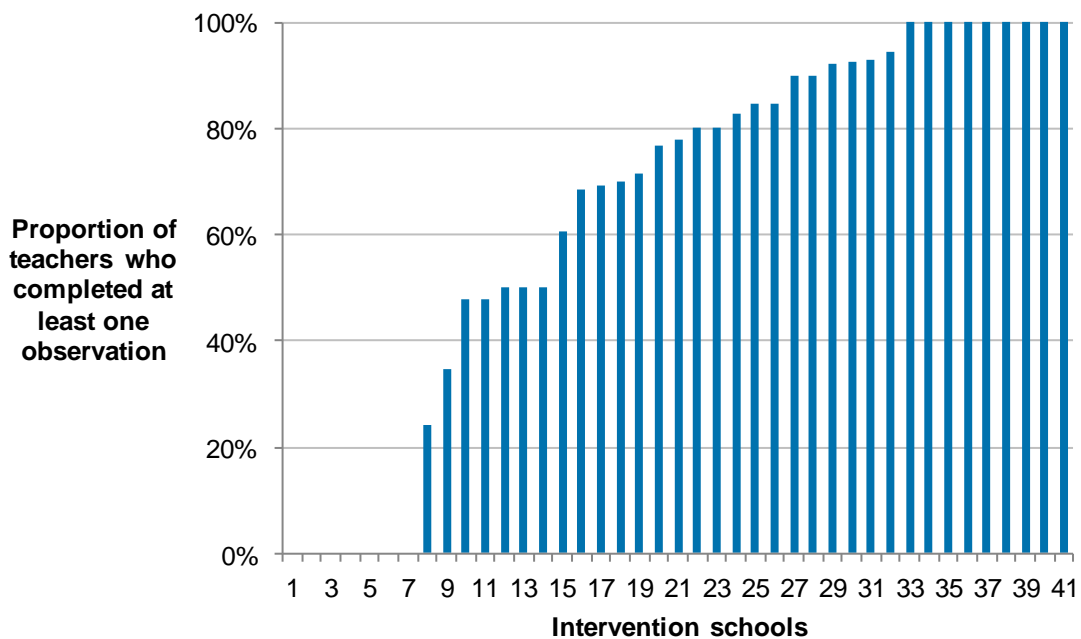
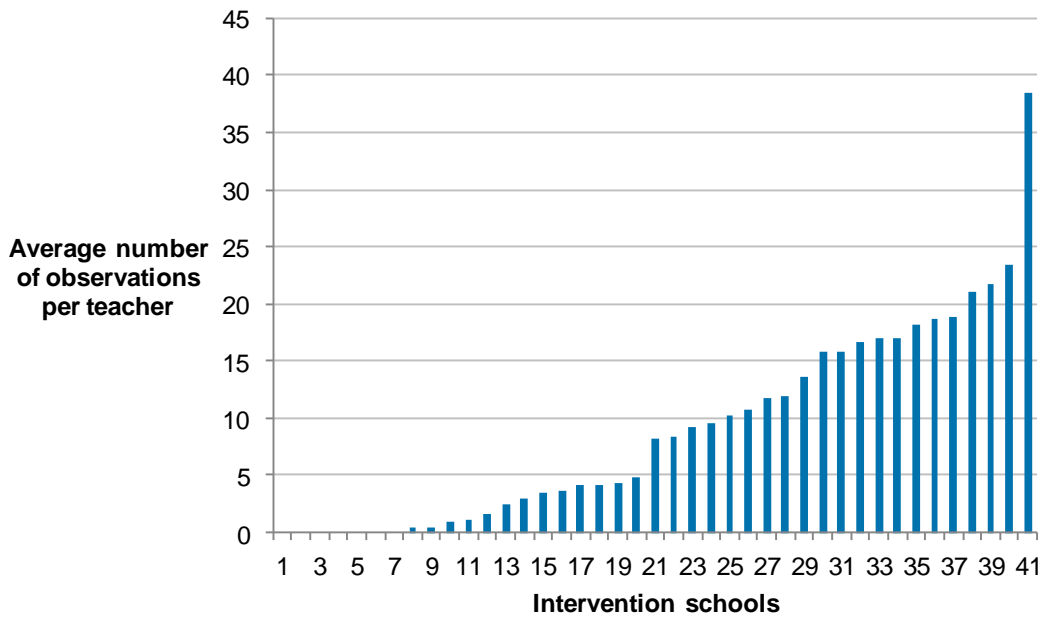


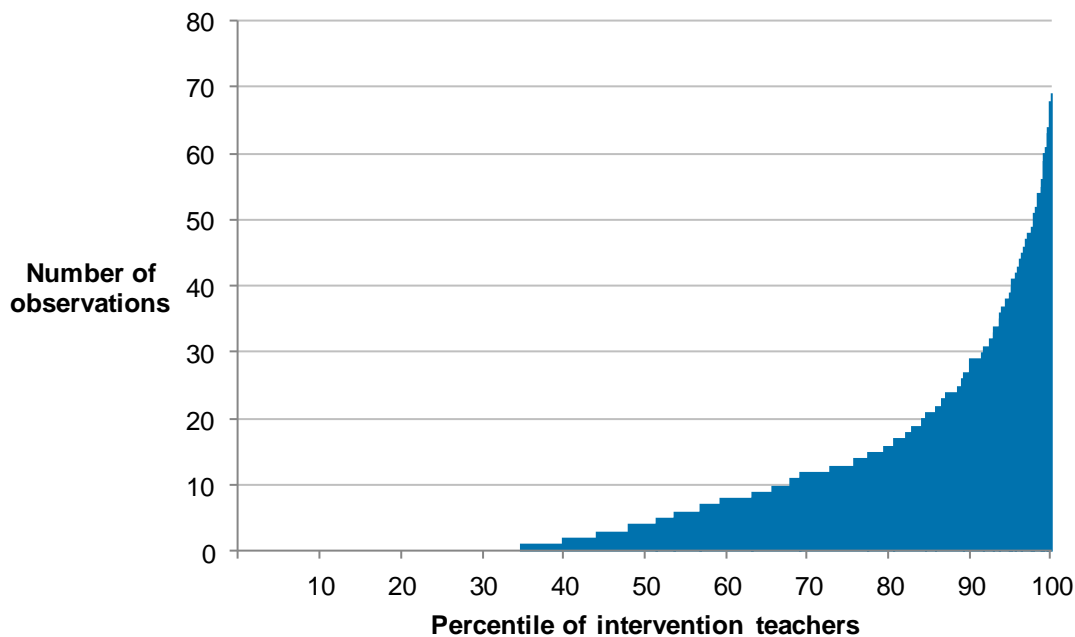
Figure 8 shows the average number of observations per teacher, which varied considerably by school. Teachers in intervention schools were involved in a total of 3,138 observations over two years, each involving two teachers—the observer and the observee. On average, teachers were involved in around ten observations each. In four intervention schools, teachers completed more than 20 observations on average, whereas 20 schools completed fewer than five observations per teacher.

Figure 8: Average number of observations per teacher, by school



Note: the data in this figure has been adjusted to avoid double-counting observations since each observation involved two teachers.

Figure 9 shows the number of observations completed by each teacher. There was a wide range: 10% of intervention teachers were involved in 29 or more observations, with five being involved in as many as 60 observations (either observing or being observed) and two almost 70. Around two thirds of teachers (65%) of the 660 intervention teachers completed at least one observation and around a third did not complete any.

Figure 9: Average number of observations per teacher over both intervention years

Note: the data in this figure double-counts observations because each observation involved two teachers.

Case study interviews indicated that teachers in low dosage departments found the number of observations in their schedules manageable and useful. Teachers in high dosage departments expressed more concerns about the manageability and sustainability of the observation programme, but even in low dosage departments some teachers felt the number of observations was demanding if they had to arrange their own cover.

'We had the low dosage. This still felt like quite a lot due to the hassle of organising cover arrangements' (maths teacher, School 2).

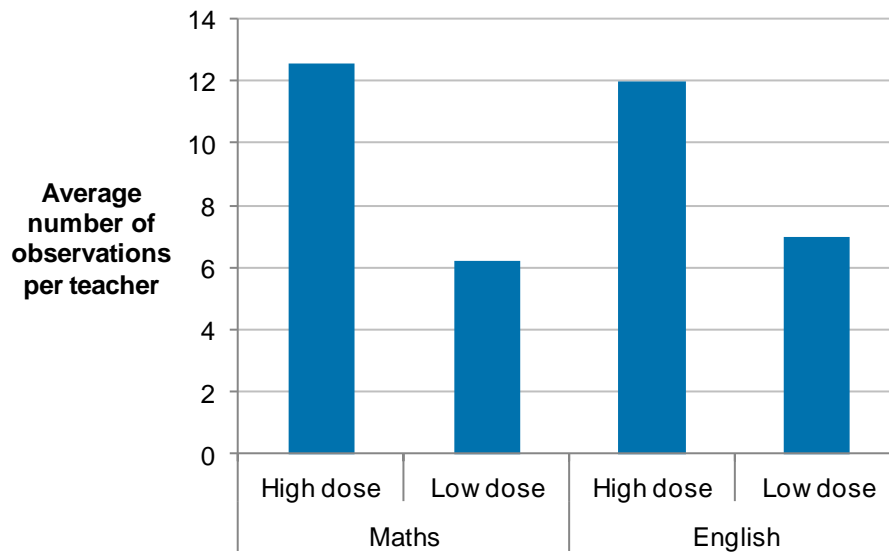
Others felt even the high dosage was manageable if properly planned:

'I think maybe one per half term would be good. It's not too much if it's timetabled in, so then you can plan for it' (English teacher, School 3).

Overall, teachers and co-ordinators tried hard to meet the dosage requirements.

Teachers appear to have been influenced by their allocated dosage. According to the RANDA TOWER data, high-dosage departments completed around twice as many observations, on average, as low-dose departments. Figure 10 shows the number of observations per teacher by department and dosage level over the two years of the evaluation. However, the level of implementation was well below the developer's initial expectations of how many observations teachers would carry out. The actual level was around a quarter of what was expected equating to six observations per year in the low-observation departments, and twelve per year in the high-observation departments (the figure measures the number of observations over two years and double-counts observations because each observation involved two teachers). The figure also shows that the level of implementation was similar, on average, between English and maths departments.

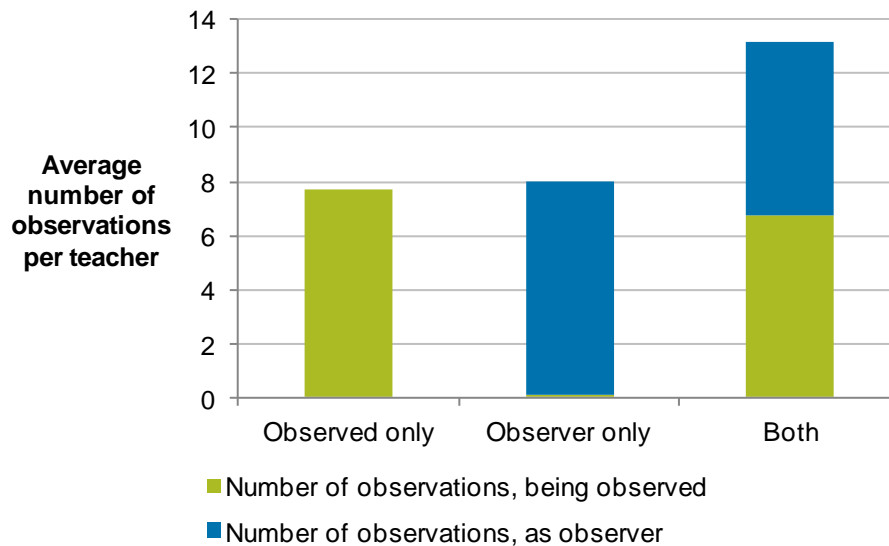
Figure 10: Average number of observations per teacher over both intervention years, by department and dosage



Note: the data in this figure double-counts observations because each observation involved two teachers.

The RANDA TOWER data also indicates that observer and observee teacher roles were almost exclusively adhered to, and confirmed that teachers allocated to both roles were involved in more completed observations than either of the restricted groups. Figure 11 shows the number of observations per teacher, by randomised role (note that the data in this figure double-counts observations because each observation involved two teachers).

Figure 11: Average number of observations per teacher over both intervention years, by randomised role



Note: the data in this figure double-counts observations because each observation involved two teachers.

Outcomes

Throughout the process evaluation, it became clear that teachers in the intervention schools were fairly evenly divided in their views. This included their general attitudes and their engagement with the programme, and their opinions about the perceived outcomes and benefits of the programme. This split may have been related to existing CPD already in the schools (or lack of it), or the enthusiasm of the head of department or project co-ordinator. Some teachers interviewed had been very pleased to use the peer-observations to actively reflect on their own practice while others felt they were being asked to carry out activities for which they could see little value.

The survey results confirmed this: around half the teachers agreed or disagreed with the majority of the attitudinal statements. For example:

- 54% said that they found involvement in the project to be of value professionally. However, 40% of respondents disagreed or strongly disagreed with this statement.
- 52% did not agree that the observation programme had improved collaboration between teachers, whereas just under half believed that it had.
- 46% of respondents felt the intervention would not improve teaching and learning across the school, while 39% thought it would.
- 44% believed that the whole school would not benefit from involvement in the intervention, but 36% thought it would.
- 40% of observees agreed or strongly agreed that 'being observed has improved my practice'.
- 45% of respondents said they did not want to continue to use the observation programme.
- 47% of respondents said they would recommend the programme to other teachers.

The survey results demonstrate that teachers had very mixed views concerning the value of the peer observations required by the intervention, and this was clearly reflected in the case studies.

It seems likely that teacher 'buy-in' and engagement would play a significant role in the successful implementation of an intervention of this kind.

Teachers interviewed during the case study visits also exhibited a wide variety of views on the impact of the Teacher Observation intervention. Some teachers felt total commitment and faith that the observations and reflection engendered by the intervention would enhance and develop practice across the school:

'It will help us develop good or better teaching and let observees as well as the observers gain strategies to develop and improve their teaching' (school co-ordinator, School 4).

'Teachers within the department get to coach and support each other with both pedagogical issues as well as subject specific aspects of teaching maths' (school co-ordinator, School 3).

'I observed lots of great strategies, many from teachers that are less experienced than myself. I am planning to incorporate more group work into my lessons next year, especially with my current Year 10 class. Good practice is going to be shared within the department and targets are set for future observations' (maths teacher, School 1).

'It fitted in very nicely with one of our objectives as a department which was to do peer observation and that was one of the things we said we wanted to do and this kind of

formalised that intent. It was certainly one of our department improvement plan points so I was glad to have it because it meant that it got done, because sometimes these things are hard to manage (English teacher, School 4).

Other teachers gave more critical perspectives where the scheduled observations were regarded simply as a hoop to jump through offering little value to practitioners or students:

'There we no benefits or positive impacts—in fact it caused disruption' (maths teacher, School 6).

'I'm not sure if it had any more impact than a normal observation. Its strengths were that it is simple, its challenges were the timetables' (English teacher, School 4).

'I feel there will be no outcomes and impacts for teaching strategies. I don't feel I have developed any new skills as a result of being involved in the programme' (maths teacher, School 2).

In the case study interviews, however, teachers allocated as observers agreed, almost unanimously, that watching other professionals at work made them reflect on their own practice, and the majority of observees were at ease with colleagues observing them.

Year 2 pro forma returns indicated that the main perceived benefits of being involved in the Teacher Observation study stemmed from teachers visiting one another's classrooms. It was felt that such visits encouraged joint reflection, the sharing good practice, and allowed teachers to draw on, and learn, from the skills and experiences of colleagues in a different subject area.

Further sub-group analysis of the views of those responding to the survey in the high dosage (n = 64) and low dosage (n = 40) groups revealed additional insights, although some caution should be taken in interpreting the responses due the small sample sizes involved.

Participants in the high dosage group were considerably more likely to report that being involved in the project had made them reflect on their own practice (75% for high dosage; 54% for low dosage). However, this difference did not appear to translate into self-perceived changes in practice, with the groups much closer in the proportions reporting that as a result of the programme they had made improvements in their day-to-day practice (44% for high dosage; 38% for low dosage).

Other benefits reported at the end of the programme included use of the observation framework to help structure and standardise classroom observations, and the two-way benefits experienced by teacher-observers who were able to both support their colleagues and reflect more deeply on their own practice.

Perception of benefit to students

The pattern was slightly different when teachers were asked whether they felt that students benefitted from the observations: 56% of respondents felt that students had not benefitted whereas only 27% felt that they had (the other 16% said they didn't know). This suggests that even where teachers felt they may have reflected more on their practice, they did not link this to positive pupil outcomes.

These mixed views were also seen in the case study interviews:

'I anticipate there will be no significant outcomes and impacts for pupils as a result of the peer observation programme' (English teacher, School 3).

'People found the idea interesting but didn't see the programme as being particularly valuable for pupils' (maths teacher, School 6).

'The value for learners depends if the teacher acts on feedback given' (maths teacher, School 1).

Unintended outcomes

In terms of potential, unintended outcomes, a number of teachers interviewed during the case study visits raised some concerns that taking time out from their teaching to conduct the observations was actually more likely to be detrimental to their students than beneficial. They believed that both they and their students were losing valuable teaching and learning time in the classroom during a crucial period of their students' preparation for GCSEs and suggested that the intervention of peer-observation would be less stressful, and of more value, if it was carried out in Year 9 classes. This observation was voiced by teachers in both high and low dosage departments.

'We should not have to leave exam classes at key points in the year' (maths teacher, School 6).

'The number of observations required was often seen as too onerous—teachers were reluctant to leave their own exam classes to carry out observations' (project co-ordinator, School 2).

'The lead-up to the exam GCSE and A-level things became quite stressful for everybody so it was putting a bit of pressure on those being observed that we had to put this in place. They had certain things they needed to get done' (project co-ordinator, School 1).

In one case study school, however, one newly qualified teacher (NQT) only taught one (low ability) GCSE class in a high dosage department. She was randomised to be an observee and was upset that, as she had only one class, other members of the department appeared in almost every lesson. She perceived this to be detrimental to her students who were distracted by the observers. This may be an example of an in-school situation that would be organised differently without the rigid requirements of a trial.

Formative findings

The case studies highlighted the importance of the in-school project co-ordinator in facilitating successful implementation of the programme. The co-ordinator appears to have an important impact on teachers' attitudes and willingness to become (and stay) involved in the Teacher Observation programme. The way in which the intervention is introduced to colleagues seems to be an important factor—sharing of information and discussing the purpose of the observations and what participants hope to gain from them. The level of support and encouragement the co-ordinator offers, particularly in terms of timetabling and cover, was also regarded by teachers as an important enabling factor. The use made of the observations (or lack of it) was a significant factor that related to how teachers perceived the value of the intervention.

Teachers made specific comments when completing the Year 2 pro formas about how the programme could be improved. The majority related to the RANDA app, or referred to how the data could be used more effectively. Suggestions included:

- giving schools the ability to edit the text within the RANDA app so that it could be tailored to individual schools' visions and ethos;

- configuring RANDA so that at the end of an observation an email is automatically sent to the teacher who has been observed, notifying them that feedback is available;
- making the notes in the RANDA app downloadable in PDF format;
- giving teachers who have been observed access to comments made by the observer; and
- making every teacher both an observee and an observer to make it easier for schools to ensure they had the capacity to complete the required number of observations.

It is possible that some of the features requested are already available within the existing software. However, users were not always aware of the ways in which the observation data could be used and some schools made no use of it at all. It seems likely that some of the barriers (timetabling and staffing issues, issues with technology and teacher engagement) may not be viewed as problematic if a school had chosen to use this kind of teacher observation as part of a whole-school CPD strategy.

Control group activity

Overall, almost three-quarters (71%) of the control group schools were already completing some peer observation prior to the intervention (Table 20).

When the control group teachers were asked about current practice in their schools (Table 23), almost 80% had been observed for performance management purposes, 72% had been observed within their department for CPD purposes, and just over half (53%) had conducted observations within their department for CPD purposes. Two-fifths of control group teachers had also been observed, or observed others, across departments (40% and 39% respectively).

As with the intervention group, having observers in class was a familiar phenomenon.

Table 23: Current peer observation practice among control group teachers

	I have been observed (%)	I have conducted observations (%)	Only senior management have conducted observations (%)	N/A (%)
Peer observations <i>within</i> departments for CPD purposes.	72	53	8	10
Peer observations <i>across</i> departments for CPD purposes.	40	39	18	27
Open lessons for CPD purposes.*	34	25	10	42
Observations for performance management purposes.	80	37	16	6

Respondents could tick more than one box therefore neither rows nor columns add up to 100%.

** For example, when teachers are invited to visit specific lessons delivered by colleagues.*

Teachers and senior leaders in both intervention and control group schools were asked in the online survey if there had been any changes to their approaches to teaching maths or English in the previous year. In response, 30% of teachers in control group schools said their schools had introduced a new

approach in English, and 40% had introduced a new approach in maths (in comparison to 17% and 22% in the intervention group, respectively).

In addition, 66% of teachers in control group schools said that something had happened in their school in the previous year that might have affected students' skills in English, and 57% said the same for maths (in comparison to 44% and 45% respectively in the intervention group). Examples given included the introduction of new curricular materials (such as resources designed with a mastery approach for maths), a change of exam board or specification for GCSEs, focused (subject-level) CPD, and targeted support programmes (such as accelerated reader).

It is possible that (other) efforts to improve teaching approaches may have been reduced slightly in intervention schools because of the Teacher Observation project. It also seems likely that involvement in the project had been regarded as an improvement initiative by senior leaders in intervention schools.

Conclusion

Key conclusions

1. The project found no evidence that Teacher Observation improves combined GCSE English and maths scores.
2. The project found no evidence of impact of the intervention on the GCSE English and maths attainment of pupils who have ever been eligible for FSM.
3. In general, teachers delivered the minimum number of observations allowed rather than the higher number suggested by the developers: teachers had difficulty fitting in the required number of observations because of timetabling and arranging cover, and some experienced problems using the software. Even when observations did take place, there was no evidence that schools which did more observations had better pupil results.
4. Teacher engagement with the programme varied greatly across schools, and practice ranged from individuals simply recording some observations using the RANDA software to whole-school, collaborative planning, discussion and reflection as part of an integrated CPD programme.
5. Almost three-quarters of the control group schools were already doing some peer observation prior to the intervention. The lack of impact seen in this study may be because the structured Teacher Observation intervention was no more effective than existing practice rather than because general peer observation has no impact.

Interpretation

The evidence from our impact evaluation suggests the Teacher Observation intervention had no impact on pupils' GCSE English and maths attainment. This result is highly secure given the large number of schools and pupils involved in the trial and low rate of attrition, most of which cannot be biased. In addition, the school-randomised design means there are few threats to internal validity.

Data on the number of observations that each teacher was involved in shows there was considerable variation in the level of implementation across teachers and schools. While seven schools did not complete any observations, more than half of teachers in 30 out of the 41 intervention schools were involved in observations to some extent. On average, teachers completed around three 'observation events' per year (involving two teachers each) in low-dosage departments and six per year in high-dosage departments. The level of implementation was therefore around a quarter of the developer's initial expectations that there would be six observations per teacher per year in the low-observation departments and twelve per teacher per year in the high-observation departments. The process evaluation found that many teachers had difficulty fitting in the required number of observations (because of timetabling and arranging cover) and some experienced problems using the software.

Incomplete implementation may therefore contribute to explaining why the evaluation found the intervention had no impact on pupil outcomes. However, our on-treatment analysis, which explored the association between teachers being involved in more observations and the outcomes of the pupils they taught, did not show a relationship between more observations and improved pupil outcomes. This provides tentative and non-causal evidence that teachers being involved in more observations does not have an overall positive impact on pupils.

Another way of looking at the result is with due regard to the time taken by teachers to carry out the observations. These tended to occur in non-contact time and occasionally required a teacher to leave their normal lesson, which required cover to be arranged. Either way, the teacher making the observation is likely doing so at the expense of another activity such as preparing lessons, marking books or, in some cases, actually teaching. A number of teachers felt that taking time out from teaching their exam classes to observe in another class was in fact detrimental to their pupils' learning. It could

be that any improvements in teaching practice, and subsequent pupil outcomes, are offset by the effect of reducing other activities.

While the *quantity* of observations seemed to be unrelated to the intervention's impact, the *quality* of observations might have had an influence. Previous research suggests the pairing of teacher observation with ongoing, school-based professional development is important for successful implementation (Shaha *et al.*, 2015). The process evaluation found that while some schools adopted the RANDA TOWER observation schedule as part of their ongoing CPD programmes and scheduled a number of additional planning, feedback and reflection meetings to support effective use of the software provided, in others, teachers conducted the observations but made no formal use of the materials beyond that. The project appeared to run most successfully when participants viewed it as peer-driven CPD.

Another possible explanation of the intervention having had no impact on pupil attainment is that the mechanism through which the intervention was intended to impact on pupil outcomes was operating on a timescale that extended beyond the evaluation. The intervention was intended to encourage teachers to reflect on their own practice, leading to improved teacher effectiveness and thereby to improved outcomes for learners. It may have been unrealistic to expect this mechanism to have an impact on pupil attainment in less than two years.

Limitations

As noted above, the results from the school- and department-level experiments have high security given the large number of schools and pupils involved in the trial and low rate of attrition, most of which cannot be biased. The school-randomised design means there are few threats to internal validity. The primary outcome variable was externally-marked GCSE examinations; this reduces the possibility that the outcome measures could be biased (compared to teacher-administered tests).

The multiplicity of secondary analyses exploring the impact on different cohorts, individual subjects, and on Year 10 test outcomes, should be interpreted with great caution. In particular, individual results should not be seen in isolation. We conducted 24 primary and secondary analyses across the three experiments and found one result which was statistically significant at the 5% level. However, as no consistent pattern in effect sizes was evident across the analyses, the positive finding is not compelling evidence that the intervention had an impact in that particular case. Under the assumption that there is no underlying 'true' effect, there is a 71% probability of one or more of the 24 analyses showing a significant effect at the 5% level.

The teacher-level experiment had a much lower level of statistical precision than was anticipated at the project outset. This was primarily caused by a higher than expected intra-cluster correlation¹⁴ driven by widespread setting of pupils by ability, particularly in maths. Future evaluation designs that consider class- or teacher-level randomisation in secondary schools should carefully consider the implications that setting might have on the precision of the final analysis, after accounting for clustering.

The sample was drawn from the schools with the highest proportions of pupils eligible for free school meals, so the results are not necessarily generalizable to all secondary schools in England.

Future research and publications

Although our analysis suggests the Teacher Observation intervention had no impact on pupil attainment, the emerging evidence base from the U.S. suggests that effective teacher observation can improve teacher effectiveness and raise student attainment. There are, however, several notable differences between some of the approaches that have been found to be effective in the U.S. and the

¹⁴ Maths: school ICC = 0.06, teacher ICC = 0.43. English: school ICC = 0.04, teacher ICC = 0.29.

Teacher Observation initiative as it is currently designed. These include the absence of features that previous research suggests are important, such as:

- the pairing of teacher observation with ongoing, school-based professional development;
- the use of outside observers; and
- annual performance bonuses for observees based on a combination of teacher value added to student achievement and observations of their classroom teaching.

The developers may wish to consider whether the features above could be integrated into the Teacher Observation intervention, but as a minimum we think ongoing, school-based professional development should be offered alongside teacher observation.

Our analysis of the impact of the intervention on attainment measured the average effect of schools' implementation of the intervention. Because of the department- and teacher-level experiments, some teachers were conducting fewer observations than others by design. All else being equal, a school-level randomised design with maximum implementation encouraged across all intervention schools might have increased the chances of detecting an impact. However, as we saw no impact when exploring the link between dosage and outcome, it is likely that such a design would also yield a null result. Either way, where interventions have limited prior evidence of their efficacy, additional conditions that weaken the intervention's dosage should be considered cautiously.

References

- Bill and Melinda Gates Foundation (2013) 'Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project's Three-Year Study', MET Project, Policy and Practice Brief, available: <http://www.edweek.org/media/17teach-met1.pdf> [21 March, 2017].
- BMJ (2014) 'Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide', *The BMJ*, 348 (7 March, 2014), available: <http://www.bmj.com/content/348/bmj.g1687> [22 March, 2017].
- Cordingley, P., Higgins, S., Greany, T., Buckler, N., Coles-Jordan, D., Crisp, B., Saunders, L. and Coe, R. (2015) 'Developing Great Teaching: Lessons from the International Reviews into Effective Professional Development', London: Teacher Development Trust, available: <http://tdtrust.org/wp-content/uploads/2015/10/DGT-Full-report.pdf> [21 March, 2017].
- Department for Education (DfE) (2011) *Teachers' Standards: Guidance for School Leaders, School Staff and Governing Bodies*, London: DfE, available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/301107/Teachers__Standards.pdf [22 March, 2017].
- Garrett, R. and Steinberg, M. P. (2015) 'Examining teacher effectiveness using classroom observation scores: evidence from the randomization of teachers to students', *Educational Evaluation and Policy Analysis*, 37 (2), pp. 224–242.
- Glazerman, S. and Seifullah, A. (2012) 'An Evaluation of the Chicago Teacher Advancement Program (Chicago TAP) After Four Years', Washington, DC: Mathematica Policy Research.
- Jensen, B. and Reichl, J. (2011) 'Better Teacher Appraisal and Feedback: Improving Performance', Victoria: Grattan Institute, available: https://grattan.edu.au/wp-content/uploads/2014/04/081_report_teacher_appraisal.pdf [21 March, 2017].
- Manzeske, D. P., Eno, J. P., Stonehill, R. M., Cumming, J. M. and MacGillivray, H. L. (2014) 'Assessing teacher effectiveness through dual-rater classroom observations: researchers and district staff partnering to create calibrated performance evaluations', paper at SREE Fall 2014 Conference 'Common Ground for Practice and Research: Targeted Improvement Initiatives', Washington, DC. 4–6 September, available: <https://www.sree.org/conferences/2014f/program/downloads/abstracts/1326.pdf> [21 March, 2016].
- Mourshed, M., Chijioke, C. and Barber, M. (2010) *How the World's Most Improved School Systems Keep Getting Better*, New York: McKinsey & Company, available: <http://www.mckinsey.com/industries/social-sector/our-insights/how-the-worlds-most-improved-school-systems-keep-getting-better> [21 March, 2017].
- O'Leary, M. and Brooks, V. (2013) 'Raising the stakes: classroom observation in the further education sector in England', *Professional Development in Education*, 40 (4), pp. 530–545.
- RANDA (2012) *RANDA Teacher Observation Tools Neutral and Flexible*, available: <http://randasolutions.com/press/randa-teacher-observation-tools-neutral-and-flexible/> [22 March, 2017].
- Richards, C. (2014) 'Judging the quality of teaching in lessons: some thoughts prompted by Ofsted's subsidiary guidance on teaching style', *FORUM*, 56 (2), pp. 199–206.

Shaha, S. H., Glassett, K. F. and Copas, A. (2015) 'The impact of teacher observations with coordinated professional development on student performance: A 27-state program evaluation', *Journal of College Teaching & Learning*, 12 (1), pp. 55–64.

Slater, H, Davies, N. and Burgess, S. (2009) 'Do teachers matter? Measuring the variation in teacher effectiveness in England', CMPO Working Paper 09/212, available: <http://www.bristol.ac.uk/media-library/sites/cmppo/migrated/documents/wp212.pdf> [03 April 2017].

Styles, B. (2014) *Teachers Observation Intervention. Protocol for Evaluation of the Teacher Observation Intervention*, Slough: NFER, available: https://v1.educationendowmentfoundation.org.uk/uploads/pdf/Round_5_-_Teacher_Observation_Intervention.pdf [22 March, 2017].

Taylor, E. S. and Tyler, J. H. (2012) 'Can teacher evaluation improve teaching? Evidence of systematic growth in the effectiveness of midcareer teachers,' *EducationNext*, 4 (12), pp. 78–84, available: <http://educationnext.org/can-teacher-evaluation-improve-teaching/> [24 March, 2017].

Office for Standards in Education, Children's Services and Skills (Ofsted) (2016) *School Inspection Handbook*, Manchester: Ofsted, available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/553940/School_inspection_handbook-section_5.doc [29 March, 2017].

TNTP (2012) 'MET' Made Simple: Building Research-Based Teacher Evaluations' (Issue Analysis Report), New York: TNTP, available: https://tntp.org/assets/documents/TNTP_METMadeSimple_2012.pdf [21 March, 2017].




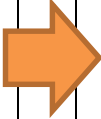


White, T. (2014) 'Adding Eyes: The Rise, Rewards, and Risks of Multi-Rater Teacher Observation Systems' (Issue Brief), California: Carnegie Foundation for the Advancement of Teaching, available: https://www.carnegiefoundation.org/wp-content/uploads/2014/12/BRIEF_Multi-rater_evaluation_Dec2014.pdf [21 March, 2017].

Appendix A: EEF cost rating

Cost ratings are based on the approximate cost per pupil per year of implementing the intervention over three years. More information about the EEF's approach to cost evaluation can be found [here](#). Cost ratings are awarded as follows:

Cost rating	Description
£ £ £ £ £	<i>Very low</i> : less than £80 per pupil per year.
£ £ £ £ £	<i>Low</i> : up to about £200 per pupil per year.
£ £ £ £ £	<i>Moderate</i> : up to about £700 per pupil per year.
£ £ £ £ £	<i>High</i> : up to £1,200 per pupil per year.
£ £ £ £ £	<i>Very high</i> : over £1,200 per pupil per year.

Appendix B: Security classification of trial findings

Rating	Criteria for rating			Initial score	Adjust	Final score
	Design	Power	Attrition¹⁵			
5 	Well conducted experimental design with appropriate analysis	MDES < 0.2	0-10%	5		5
4 	Fair and clear quasi-experimental design for comparison (e.g. RDD) with appropriate analysis, or experimental design with minor concerns about validity	MDES < 0.3	11-20%		Adjustment for Balance [0]	
3 	Well-matched comparison (using propensity score matching, or similar) or experimental design with moderate concerns about validity	MDES < 0.4	21-30%			
2 	Weakly matched comparison or experimental design with major flaws	MDES < 0.5	31-40%		Adjustment for threats to internal validity [0]	
1 	Comparison group with poor or no matching (E.g. volunteer versus others)	MDES < 0.6	41-50%			
0 	No comparator	MDES > 0.6	over 50%			

- **Initial padlock score:** lowest of the three ratings for design, power and attrition = 5 padlocks
- **Reason for adjustment for balance** (if made): N/A
- **Reason for adjustment for threats to validity** (if made): N/A
- **Final padlock score:** initial score adjusted for balance and internal validity = 5 padlocks

The design is a randomised controlled trial. MDES is 0.12 at randomisation, pupil level attrition is approx. 9% for the primary analyses. No threats to validity or imbalance are present.

¹⁵ Attrition should be measured at the pupil level (even for clustered trials) and from the point of randomisation to the point of analysis.

Appendix C: Observation domains and components

DOMAIN 1: THE CLASSROOM ENVIRONMENT				
Component	Ineffective (1-3)	Basic (4-6)	Effective (7-9)	Highly Effective (10-12)
1a Creating an Environment of Respect and Rapport	Classroom interactions, both between teacher and students and among students, are negative, inappropriate, or insensitive to students' cultural backgrounds, ages and developmental levels. Student interactions are characterised by sarcasm, put-downs, or conflict.	Classroom interactions, both between the teacher and students and among students, are generally appropriate and free from conflict, but may reflect occasional displays of insensitivity or lack of responsiveness to cultural or developmental differences among students.	Classroom interactions, both between teacher and students and among students, are polite and respectful, reflecting general warmth and caring, and are appropriate to the cultural and developmental differences among groups of students.	Classroom interactions, both between teacher and students and among students, are highly respectful, reflecting genuine warmth and caring and sensitivity to students' cultures and levels of development. Students themselves ensure high levels of civility among members of the class.
1b Establishing a Culture for Learning	The classroom environment conveys a negative culture for learning, characterised by low teacher commitment to the subject, low expectations for student achievement, and little or no student pride in work.	The teacher's attempts to create a culture for learning are partially successful, with little teacher commitment to the subject, modest expectations for student achievement, and little student pride in work. Both teacher and students appear to be only "going through the motions."	The classroom culture is characterised by high expectations for most students and genuine commitment to the subject by both teacher and students, with teacher demonstrating enthusiasm for the content and students demonstrating pride in their work.	High levels of student energy and teacher passion for the subject create a culture for learning in which everyone shares a belief in the importance of the subject and all students hold themselves to high standards of performance they have internalized.
1c Managing Classroom Procedures	Much teaching time is lost because of inefficient classroom routines and procedures for transitions, handling of supplies, and performance of non-teaching duties. Students not working with the teacher are not productively engaged in learning. Little evidence that students know or follow established routines.	Some teaching time is lost because classroom routines and procedures for transitions, handling of supplies, and performance of non-teaching duties are only partially effective. Students in some groups are productively engaged while unsupervised by the teacher.	Little teaching time is lost because of classroom routines and procedures for transitions, handling of supplies and performance of non-teaching duties, which occur smoothly. Group work is well-organised and most students are productively engaged while working unsupervised.	Teaching time is maximised due to seamless and efficient classroom routines and procedures. Students contribute to the seamless operation of classroom routines and procedures for transitions, handling of supplies, and performance of non-instructional duties. Students in groups assume responsibility for productivity.
1d Managing Student Behaviour	There is no evidence that standards of conduct have been established, and there is little or no teacher monitoring of student behaviour. Response to student misbehaviour is repressive or disrespectful of student dignity.	It appears that the teacher has made an effort to establish standards of conduct for students. The teacher tries, with uneven results, to monitor student behaviour and respond to student misbehaviour.	Standards of conduct appear to be clear to students, and the teacher monitors student behaviour against those standards. The teacher's response to student misbehaviour is consistent, proportionate, appropriate and respects the students' dignity.	Standards of conduct are clear, with evidence of student participation in setting them. The teacher's monitoring of student behaviour is subtle and preventive, and the teacher's response to student misbehaviour is sensitive to individual student needs and respects students' dignity. Students take an active role in monitoring the standards of behaviour.

Teacher Observation

1e Organising Physical Space	The physical environment is unsafe, or some students don't have access to learning. There is poor alignment between the physical arrangement of furniture and resources and the lesson activities.	The classroom is safe, and essential learning is accessible to most students; the teacher's use of physical resources, including computer technology, is moderately effective. The teacher may attempt to modify the physical arrangement to suit learning activities, with limited effectiveness.	The classroom is safe, and learning is accessible to all students; the teacher ensures that the physical arrangement is appropriate for the learning activities. The teacher makes effective use of physical resources, including computer technology.	The classroom is safe, and the physical environment ensures the learning of all students, including those with special needs. Students contribute to the use or adaptation of the physical environment to advance learning. Technology is used skilfully, as appropriate to the lesson.
DOMAIN 2: TEACHING				
Component	Ineffective (1-3)	Basic (4-6)	Effective (7-9)	Highly Effective (10-12)
2a Communicating with Students	Expectations for learning, directions and procedures, and explanations of content are unclear or confusing to students. The teacher's written or spoken language contains errors or is inappropriate for students' cultures or levels of development.	Expectations for learning, directions and procedures, and explanations of content are clarified after initial confusion; the teacher's written or spoken language is correct but may not be completely appropriate for students' cultures or levels of development.	Expectations for learning, directions and procedures, and explanations of content are clear to students. Communications are accurate as well as appropriate for students' cultures and levels of development. The teacher's explanation of content is scaffolded, clear, and accurate and connects with students' knowledge and experience. During the explanation of content, the teacher focuses, as appropriate, on strategies students can use when working independently and invites student intellectual engagement.	Expectations for learning, directions and procedures, and explanations of content are clear to students. The teacher links the instructional purpose of the lesson to the wider curriculum. The teacher's oral and written communication is clear and expressive, appropriate to students' cultures and levels of development, and anticipates possible student misconceptions. The teacher's explanation of content is thorough and clear, developing conceptual understanding through clear scaffolding and connecting with students' interests. Students contribute to extending the content by explaining concepts to their peers and suggesting strategies that might be used.
2b Using Questioning and Discussion Techniques	The teacher's questions are of low cognitive challenge or inappropriate, eliciting limited student participation, and recitation rather than discussion. A few students dominate the discussion.	Some of the teacher's questions elicit a thoughtful response, but most are low-level, posed in rapid succession. The teacher's attempts to engage all students in the discussion are only partially successful.	Most of the teacher's questions elicit a thoughtful response, and the teacher allows sufficient time for students to answer. All students participate in the discussion, with the teacher stepping aside when appropriate.	Questions reflect high expectations and are culturally and developmentally appropriate. Students formulate many of the high-level questions and ensure that all voices are heard.
2c Engaging Students in Learning	Activities and assignments, materials, and groupings of students are inappropriate for the learning outcomes or students' cultures or levels of understanding, resulting in little intellectual engagement. The lesson has no clearly defined structure or is poorly paced.	Activities and assignments, materials, and groupings of students are partially appropriate for the learning outcomes or students' cultures or levels of understanding, resulting in moderate intellectual engagement. The lesson has a recognisable structure but is not fully maintained and is marked by inconsistent pacing.	Activities and assignments, materials, and groupings of students are fully appropriate for the learning outcomes and students' cultures and levels of understanding. All students are engaged in work of a high level of rigour. The lesson's structure is coherent, with appropriate pace.	Students, throughout the lesson, are highly intellectually engaged in significant learning and make material contributions to the activities, student groupings, and materials. The lesson is adapted as needed to the needs of individuals, and the structure and pacing allow for student reflection and closure.

Teacher Observation

2d Use of Assessment	Assessment is not used in teaching, either through monitoring of progress by the teacher or students, or adequate feedback to students. Students are not aware of the assessment criteria used to evaluate their work, nor do they engage in self or peer-assessment.	Assessment is occasionally used in teaching, through some monitoring of progress of learning by the teacher and/or students. Feedback to students is uneven, and students are aware of only some of the assessment criteria used to evaluate their work. Students occasionally assess their own or their peers' work.	Assessment is regularly used in teaching, through self- or peer- assessment by students, monitoring of progress of learning by the teacher and/or students, and high-quality feedback to students. Students are fully aware of the assessment criteria used to evaluate their work and frequently do so.	Assessment is used in a sophisticated manner in teaching, through student involvement in establishing the assessment criteria, self-or peer assessment by students, monitoring of progress by both students and the teacher, and high-quality feedback to students from a variety of sources. Students use self-assessment and monitoring to direct their own learning.
2e Demonstrating Flexibility and Responsiveness	The teacher adheres to the lesson plan, even when a change would improve the lesson or address students' lack of interest. The teacher brushes aside student questions; when students experience difficulty, the teacher blames the students or their home environment.	The teacher attempts to modify the lesson when needed and to respond to student questions, with moderate success. The teacher accepts responsibility for student success but has only a limited repertoire of strategies to draw upon.	The teacher promotes the successful learning of all students, making adjustments as needed to lesson plans and accommodating student questions, needs, and interests.	The teacher seizes an opportunity to enhance learning, building on a spontaneous event or student interests, or successfully adjusts and differentiates instruction to address individual student misunderstandings. The teacher ensures the success of all students by using an extensive repertoire of teaching strategies and soliciting additional resources from the school or community.

Appendix D: Feasibility study findings

This appendix presents the findings of the initial feasibility study, as reported to EEF in June 2014. The report was authored by NFER researchers Dr Anneka Dawson and Juliet Sizmur. The feasibility study was known at the time as the 'pilot', but had a more limited scope than an EEF pilot evaluation.

Pilot Summary Report

Introduction

The teacher observation intervention is being delivered by the Centre for Market and Public Organisation (CMPO), 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 frequent peer observations (teachers observing and being observed) over the course of a year in maths and English departments in 120 secondary schools. Teachers are using a tablet with RANDA software to record the observations. The National Foundation for Educational Research (NFER) is carrying out an independent evaluation of the intervention.

CMPO piloted the research design in three schools in March/April 2014. The purpose of the pilot was 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 to interview senior leaders and those taking part in the pilot - **NFER**
4. test the design and randomisation - whether the split across maths and English departments works; whether the learner-/teacher-level information is correct in respect of 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**.

This summary report discusses NFER's contribution to the pilot and therefore focuses on items 3 and 5 above. As CMPO did not want to ask the pilot schools for class and teachers lists, we were unable to test the design and randomisation. As a result, this remains a potential risk for the implementation of the main trial, as different levels of both dosage (department level randomisation) and role (observer/observee/both as part of teacher level randomisation) will need to work in each school.

Process evaluation

We visited one school to observe the introduction of the materials, then visited all three pilot schools to conduct follow-up interviews with senior leaders and, where possible, participating teachers. We interviewed four senior leaders and seven teachers. All schools were able to deliver the pilot as planned. All interviewees agreed that advance planning/timetabling of the observations would be critical for the main study to work properly, especially if all colleagues in the two departments were to be involved. Cover supervisors would need to be allocated for at least some of the observations, and there would be associated costs. The interviewees expressed some concern about what to do with teachers who don't 'buy in'.

A summary of the main findings is outlined below.

- Overall teachers reported positive experiences of using the software. They felt it was straightforward to learn and intuitive to use.
- There were some initial problems with log-in, user IDs and the initial 'opening' of the programme but, after these issues were addressed/clarified, the software worked well.
- All three schools provided cover for the pilot, but all agreed that they would expect teachers to use their free periods if the programme had to be maintained over a longer period.
- All interviewees agreed that the observation tool would be valuable for continuing professional development (CPD), although the extent to which time was allocated for planning and feedback among peers varied considerably.
- Some senior leaders felt the observations could also be used to feed into performance management discussions; others felt they would use anonymised data to feed into their departmental reviews.
- Senior leaders felt that teachers would need some more specific training on observation and giving feedback.
- Teachers felt they gained more from observing than from being observed.
- Most interviewees felt that all teachers in a department should be involved; some thought only selected teachers.
- Senior leaders thought timetabling the observations across all teachers in a department would be a major exercise, which would need to be planned early.
- All schools commended the helpfulness of CMPO staff.

The software and the rubric

Interviewees were positive about the software and the rubric.

- All interviewees agreed that it was fairly intuitive, but said that they would want more time to familiarise themselves with the categories and agree priorities/areas of focus beforehand.
- After being introduced to the software by a member of the project team, teachers reported spending from five minutes to three hours familiarising themselves with the programme before beginning their observations.

- All interviewees preferred using the tick-boxes to the 'free-form' recording of observations, but liked having the different options available.
- They thought the rubrics were easy to apply and appropriate for both subject areas.
- One school laminated the rubric so that observers could study it before the observations.
- There were some concerns about ensuring a shared understanding of what each rating means.
- Interviewees expressed some reservations about 'rating' peers – most felt that formative feedback fitted better with collaborative endeavour.
- Some felt the rubrics were very long (and in places repetitive, and could/should be streamlined - for example some areas could be combined). One teacher suggested a maximum of six areas to assess – e.g. three on instruction and three on classroom procedures.
- Both box-ticking and free-form observations could take place in lesson time (and not beyond) which interviewees felt is a benefit of the system.
- All teachers and senior staff believed that the follow-up discussions were of great importance for improving teaching and learning across the school (identifying elements of good practice to be shared and to encourage group and individual reflection and coaching).

Number of observations

Interviewees in the pilot expressed concern about the number of observations and the cover required and suggested that careful planning will be needed to integrate the programme.

- Almost all interviewees agreed that to do six observation periods per year would be sustainable. One felt that four each year would be a more realistic maximum. However, most agreed that it would be possible to do two observations per school period.
- Those interviewed generally felt that more than six observations in a given year would be difficult to sustain, both in terms of cover and in terms of value/thoroughness. All agreed that in order to get the best out of the observation practice they would need to set aside additional time for joint planning and feedback (and time for self-reflection too).
- In some schools, all English/maths lessons for one year group took place at the same time. Cover would therefore need to be provided and senior leaders considered this to be a substantial commitment, which would have to be offset against the perceived value of participating in the programme. Alternatively, teachers felt there would still be value in observing the teaching in other year groups, or across subjects.
- Teachers were very enthusiastic about observing teaching styles in other departments, but felt that strong subject knowledge was essential if they were to give useful feedback.

Views on effectiveness of the programme

Interviewees enjoyed taking part in the pilot and felt that if it was rolled out it would have positive potential impacts on all stakeholders.

- All teachers and senior leaders felt that they, and their students, would benefit from their involvement in an observation programme like the Teacher Observation Intervention.

- They generally agreed that it should be emphasised that the programme should be presented and used as a collaborative development tool and not to make judgements. They also agreed that the programme could change the culture of observation practice.
- Interviewees were also of the opinion that the programme would be good for developing reflective practitioners; for collaborative practice and planning; for subject field development; to identify areas of strength and potential weakness in departments; and to inform training needs.
- All reported that they would recommend the peer observation programme to other schools.
- Interviewees commented that they felt the programme would help improve teaching and learning and would encourage teachers to develop their own practice. However, some felt that teachers will need considerable support to change.

Recommendations

Based on the findings from the process evaluation interviews we suggest that the following changes could be implemented in time for the main trial in September 2014:

- **Reverse the ratings system:** The 1-4 ratings in the software were counter-intuitive as schools are used to Ofsted ratings where 1 is 'outstanding' and 4 'inadequate'. Teachers felt it would be easier to make their observations if the 1-4 ratings in the programme reflected this.
- **Simplify utilisation of the observation reports:** Some teachers said they would like to be able to download/email/print documents from the tablets for their records and to use in discussion during feedback sessions. At the moment they felt that accessing the information was rather 'clunky'. NB. It may already be possible to do this – if so, it may be worth highlighting this during the training.
- **Anglicise the rubric language:** Some of the language and terminology is very American (e.g. 'instruction' instead of 'learning'). It would be good if this could be adapted.

In addition, the following issues should be taken into consideration for any future roll-out of the programme:

- Some schools would like to be able to customise their own rubrics, but accept that this may be better after the main study.
- Timetabling can be extremely complex in schools and organising cover can be difficult. Schools therefore need maximum time to plan and implement a programme like this on a large scale. They would benefit from knowing they are taking part in a programme like the Teacher Observation Intervention in the summer term (if not earlier) of the school year before the programme takes place. This would enable them to plan for the programme more easily and extra staff can be employed as cover supervisors/ temporary staff if necessary.

- Teachers would benefit from more thorough observation training so that a consistent standard is reached and this would be more meaningful if it was within a whole-school approach to observation.

Pilot tests

Four tests (two maths tests and two reading tests) have been constructed and will be trialled in the pilot schools before the end of the summer term 2014.

For each of the maths and reading tests, a panel of experts was convened to review existing Key Stage 3 tests (over ten years for maths, seven years for reading) and all test items were rated in terms of the following criteria:

- appropriate reflection of the GCSE subject and assessment objectives
- demand in terms of higher order thinking skills
- estimated level of difficulty for Year 10 students.

A series of item selection workshops then took place to identify items that covered a range of assessment focuses across the range of available items from different years of publication.

The aim is to produce two, one-hour tests for use in Year 10.

At this stage, we will pilot 200% of test items and conduct a detailed item analysis so that the items selected for the main study will be of a suitable level of difficulty and offer the best possible discrimination between students, in terms of general mathematical and reading skills and their application.

The tests will be marked and analysed over the summer months with feedback to schools in September 2014.

Appendix E: Memorandum of Understanding

Agreement to participate in the Evaluation of Teacher Peer Observation Project

Please sign both copies, retaining one and returning the second copy to Julia Carey at CMPO, University of Bristol.

School Name: _____

Aims of the Evaluation

The aim of this project is to evaluate the impact of the Teacher Peer Observation Project on GCSE and Year 10 test outcomes in Mathematics and English. The results of the research will contribute to our understanding of what works in raising the pupil's attainment and will be widely disseminated to schools in England. We hope that the project will make teacher observation more useful and support teacher effectiveness.

The Project

This two-year project is for teachers teaching Maths or English in GCSE classes. This is peer observation, so other teachers in your own school who teach GCSE English or GCSE Maths will conduct each observation. The observer has an iPad (which we provide) and the rubric (framework on which the observations will be based) is loaded on to this as an app. The observer watches the lesson and works through the questions. This has been designed based on extensive and robust research on lesson observation with the additional benefit of the project team's experience in this field. Furthermore advisors with substantial expertise in this regard have been engaged in designing the system.

More detail on the project is contained in the Appendix to this Agreement.

Structure of the Evaluation

The project will be evaluated using a randomised control trial: of all the schools which sign up, some will be randomly chosen to participate and others will act as control schools. The outcomes are GCSE scores in English and Maths, and specially-designed Year 10 tests that NFER will conduct for us. We will also conduct an online survey of teachers taking part and you may be asked to take part in a case study visit.

The project is being undertaken by CMPO, in association with NFER. It is funded by the Education Endowment Foundation.

Random allocation is essential to the evaluation as it is the best way of understanding the effect of Teacher Peer Observation on children's attainment. It is important that schools understand and consent to this process.

Use of Data

All data on teachers and pupils will be treated with the strictest confidence. Teachers will be anonymised for the purposes of the observation, and teacher names are not required. Pupils

will also be anonymous and we will require you to provide us with UPNs no later than Friday, 11th July 2014 to match to the National Pupil Database to later retrieve GCSE outcomes, prior test scores and pupil demographics. The Year 10 tests will be collected and assessed by NFER, again just using the UPNs to link to pupils. No individual school or pupil will be identified in any report arising from the research.

We take the security of personal data very seriously. The University of Bristol and NFER are compliant with all the relevant regulations; our systems have been developed and refined over a number of years of working with personal data; and the data storage framework for this project has been checked and authorised by the University's Data Protection Officer and NFER's code of practice committee.

Responsibilities

CMPO WILL:

- Provide in-depth face-to-face training for one person per school on the operation of the software.
- Be the first point of contact for any questions about the evaluation.
- Provide on-going support to the school.
- Offer each school a charitable donation of £1000 for taking part, payable in two £500 instalments, one at the end of each academic year. At the end of the project, schools will retain the iPads allocated to them at the start of the project.
- Each half-term, we will provide each observed teacher with a report on the outcomes from observations carried out over that period (as long as there is more than one)
- Analyse the attainment data in relation to teacher observations.

NFER (NATIONAL FOUNDATION FOR EDUCATIONAL RESEARCH) WILL:

- Conduct the random allocation of schools to either the intervention or control group.
- Conduct the random allocation of Maths and English departments within intervention schools to either be high or low dosage of observation (see FAQs).
- Conduct the random allocation of teachers: at the start of the project, two thirds of teachers within the intervention group will be randomly chosen (electronically) to be observed among both GCSE Maths teachers and GCSE English teachers and two thirds will be observers.
- Set, manage, collect and assess the Year 10 tests.
- Provide an independent analysis of the attainment data.
- Provide an independent analysis of the peer observation programme. Provide the school with attainment data after the trial has been completed.

THE SCHOOL WILL:

- Consent to random allocation and commit to the outcome (whether treatment or control).
- Agree that Headteachers are consenting on behalf of all teachers taking part in the trial and for pupils' data to be used as follows: Pupils' test responses and all other pupil data will be treated with the strictest confidence. The Year 10 test responses will be collected by NFER and shared with CMPO. Pupil data will be matched with the National Pupil Database using the pupil UPNs and shared with CMPO, EEF, EEF's

data archive and the UK Data Archive for research purposes. No individual school, teacher or pupil will be identified in any report arising from the research.

- Allow time (about an hour) for the Year 10 tests for the English and Maths students and liaise with the evaluation team to find appropriate dates and times for testing to take place in Summer 2015 and Summer 2016
- Provide the data requested by CMPO as follows:
 - List of teachers to be observed and to be observers (anonymised as above)
 - Class lists: that is, a list of pupils (identified only by their UPNs) with the teacher (identified by an anonymous id) and subject and year. To be sent in excel format to **cm-po-top@bristol.ac.uk no later than Friday, 11th July 2014**
 - The predicted GCSE score for each pupil (UPN) for English and for Maths
 - These will need to be done twice, once for the academic year 2014/15 and once for 2015/16.
- Agree for us to conduct a short web-based survey of the teachers (we will ask you to email a weblink to each teacher)
- Agree to allow NFER to contact you regarding a case study visit to interview teachers and senior management about the project. This will only happen in six schools each year.
- Release a staff member for the training session on the observation software.
- Facilitate the implementation of the observation schedule and make the appropriate cover arrangements.
- Ensure that observations are conducted using the required software and rubric.
- Ensure the shared understanding and support of all school staff for the project.
- Be a point of contact for parents / carers seeking information on the project.
- Abide by the terms of RANDA software usage policy, noted overleaf

We commit to the terms of the Evaluation of Teacher Observation Project as detailed in this document.

HEAD TEACHER [NAME]: _____

OTHER RELEVANT SCHOOLS STAFF [NAMES]: _____

DATE: _____

RANDA Solutions Usage Policy

Access to the TOWER Application Software (the "Software") and information technology services (the "Service") shall be provided to the participating schools. The Software has been developed and is owned by R&A Solutions, Inc., d/b/a RANDA Solutions ("RANDA") and is protected under patent pending regulations and registered at PCT/US App. No. 12/32918. RANDA, hereby and through its agreement with Bristol University, provides a license to each participating school as a ("Licensee"). RANDA grants to Licensee a limited, non-exclusive, non-assignable, non-transferable, internal-use-only license (the "License") to the Software and Service. The License shall commence upon training by Bristol and log in by each school. The License shall be coterminous with the agreement between RANDA and Bristol University or terminated immediately for any violation of this Memorandum of Understanding, the Terms of Use or other policies of RANDA. Licensee acknowledges and agrees that all right, title and ownership interest to the Software, any and all future developments, inventions, improvements, and the Program and Documentation rests solely and exclusively in RANDA and its licensors, including all rights to patents, copyrights, trademarks, trade secrets and other intellectual property rights inherent therein or appurtenant thereto. All rights not expressly granted to Licensee

herein are reserved to RANDA. ALL PRODUCTS AND SERVICES OF RANDA, INCLUDING BUT NOT LIMITED TO THE PROGRAM AND THE SERVICE, ARE PROVIDED “AS IS”, WITHOUT WARRANTY OF ANY KIND, EXPRESSED OR IMPLIED, EITHER IN FACT OR BY OPERATION OF LAW, STATUTORY OR OTHERWISE, WHETHER BY APPLICATION OF THE UCC OR ICC. RANDA shall not be liable for any violation of law or regulation occasioned by the use of the Software, including, but not limited to: (i) Licensee's misuse of the Program, (ii) any personal injury or death occurring at Licensee's premises, (iii) Licensee's breach of any representation, warranty or obligation under this Memorandum of Understanding, or (iv) any content of Licensee's databases, including, without limitation, content which: (a) is false, fraudulent, inaccurate or misleading; (b) infringes any third party's copyright, patent, trademark, trade secret or other proprietary rights or rights of publicity or privacy; (c) violates any law, statute, ordinance or regulation; (d) is defamatory, trade libelous, unlawfully threatening, unlawfully harassing or obscene; or (e) contains any viruses, trojan horses, worms, time bombs, cancelbots, easter eggs or other computer programming routines that may damage, detrimentally interfere with, surreptitiously intercept or expropriate any system, data or personal information. RANDA, RANDA'S AFFILIATES AND LICENSORS SHALL HAVE NO LIABILITY WITH RESPECT TO THIS AGREEMENT OR OTHERWISE FOR SPECIAL, INCIDENTAL, INDIRECT, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES, INCLUDING DAMAGES FOR LOSS OF BUSINESS AND LOSS OF PROFITS, BUSINESS INTERRUPTION AND LOSS OR CORRUPTION OF DATA ARISING OUT OF THE USE OF OR INABILITY TO USE THE PROGRAM.

Teacher Peer Observation Project

Who will be involved?

This project is for teachers teaching Maths or English in GCSE classes. So the teachers who will be involved are those teaching GCSE English or GCSE Maths. This will not include iGCSEs or other equivalent qualifications. About half of the schools taking part in the trial will receive the intervention (the intervention group) and the other schools will act as a control group who continue with their current method of observations.

How does a lesson observation in this project actually work?

This is peer observation, so other teachers in your own school who teach GCSE English or GCSE Maths will conduct each observation. The observer has an iPad (which we provide) and the rubric (framework on which the observations will be based) is loaded on to this as an app. The observer watches the lesson and works through the rubric. The interface is extremely straightforward to use and intuitive, and largely involves selecting options rather than extensive typing. From our pilot, we estimate that this might take 15 – 20 minutes of the lesson. A confirmation that the observation has taken place is uploaded to us via the software company's website (see Data Protection statements below), along with a date/time stamp and the rubric responses. The rubric has been designed based on extensive and robust research on lesson observation with the additional benefit of the project team's experience in this field. Furthermore advisors with substantial expertise in this regard have been engaged in designing the rubric.

Will our school be compensated for the costs in providing this information?

Yes. We will offer each school a charitable donation of £1000 for this, payable in two £500 instalments, one at the end of each academic year. At the end of the project, schools will retain the iPads allocated to them at the start of the project. Hopefully, the main benefit to the school will be the enhancement to the usefulness of lesson observation, improved teaching and learning, and improved GCSE outcomes.

Will you provide training in observation and on-going support?

Yes. We will provide in-depth face-to-face training for one person per school. The software on the iPad provides extensive training for the other observers. There is a library of videos of lessons which the observer is asked to observe and respond on the rubric. This can be repeated until the observer is proficient. There is also a great deal of support available on the website of the software provider,

RANDA. Finally, the project team are conversant with the use of the software and can also provide support.

What kind of feedback information is available to the observed teacher?

Feedback is important for the development of the observed teacher. Each half-term, we will provide each observed teacher with an average score of the outcomes from observations carried out over that period (as long as there is more than one). We expect the observed teacher will have discussions with her/his observers and the rubric and supporting CPD materials (available within the software) will provide a good framework for such feedback.

How are teachers selected to be observed?

At the start of the project, two thirds of teachers will be randomly chosen (electronically) to be observed among both GCSE Maths teachers and GCSE English teachers.

Who will the observers be?

The observers will be from your school and will also be GCSE English or Maths teachers. It is important that the observers are as similar as possible to the observed teachers, so the observers cannot be from subjects other than Maths and English. They will also be selected at random (electronically) from the set of all GCSE English and Maths teachers. Two thirds of English and Maths teachers will be randomly selected to be observers, so some teachers will be both observed and observer. Observations will be undertaken by peers teaching within the same department or across departments but only from those teachers who teach GCSE Maths and English.

How often will a teacher be observed?

We want to find out whether frequency of observation matters. So one of the two departments involved will be high frequency and the other will be low frequency. We are fine tuning what these might be, but we anticipate a low frequency of around 6 – 9 times a year, and a high frequency of around 15 – 18 times a year.

How will the observations be organised?

We will set out a broad timescale, but in order not to interfere with the daily operations of the school the detailed scheduling will be up to you. Each observed teacher will have a required number of observations to be conducted within each two-week window (as set by the project team), but exactly when these happen within that window is up to the observee and observer to jointly determine. In terms of providing cover for the observer during her/his 15 – 20 minute absence, again this is at your discretion; in the pilot schools, some observations were covered by senior school staff, others by other colleagues and some by Teaching Assistants (TAs). It may also be possible for the observer to carry out two separate observations (on different teachers) during the same lesson period.

Who will see the observation results?

The observed teacher will see the observation results. Each half-term, we will provide each observed teacher with an average score of the outcomes from observations carried out over that period (as long as there is more than one). This summary information will also be available to the teacher's line manager should they request it (provided the observed teacher is agreeable).

Will teachers be anonymous in this study?

Yes. We will randomly generate a series of numbers and ask you to allocate one number to each teacher. We will ask you to indicate to us which of those have been used. We will ask our software providers to load these numbers into their database so outcomes are attached to the randomly generated number and there is no way for either the University or the software company to establish the true identity of observers or observees. However, it is essential that, once teachers are provided with their numbers, they make consistent use of them.

What information do you need about our students?

We need to know which pupils are taught by which teachers. So we need class lists for Maths and English GCSEs, linking pupils to teachers. Pupils will be identified by their UPNs, which we will later match to the National Pupil Database (NPD) for the analysis of the project's results. Teachers will be identified by the anonymous ID noted above. See also below on data security. Additionally we will be conducting short (approximately 50-60 minute) end of year tests in English and Maths for Year 10 students. These tests will be designed and administered by the National Foundation for Educational Research (NFER). Results of these tests will be made available to you at the end of the trial. These tests will cover general curriculum based skills but will not require revision on the part of the students.

What are the IT issues?

The app itself is extremely intuitive and easy to use, and the process worked very smoothly in the pilot schools. Since the school will retain the iPads after the project, it makes sense for the school IT team to set them up. We will provide instructions on how to download the app and can provide phone assistance if needed. But it is simply a case of downloading an app on to the tablet from the itunes store, just like any other. The data from the tablets are automatically uploaded, no-one needs to write reports or email files or plug them into networks.

How is the data safe-guarded?

We take the security of personal data very seriously. The University of Bristol is compliant with all the relevant regulations; our systems have been developed and refined over a number of years of working with personal data; and the data storage framework for this project has been checked and authorised by the University's Data Protection Officer. There are three levels to the data protection for this project. First, the observation software on the iPads is password protected and each individual will have only role-specific access to certain data. Second, once the data is uploaded to the app, the software company's chosen storage company, Peak10, will host the data. Peak 10 is Safe Harbour registered. Details of its registration can be viewed at this URL: <http://safeharbor.export.gov/companyinfo.aspx?id=20646>. Thirdly, once the data reaches the University of Bristol, where it will be stored, it is subject to the University's very stringent data security procedures. Details of these are available on request. Data on pupils will come from the NPD provided by DfE, and they too have very strict data protection requirements with which we are compliant. Data will be shared with NFER and then transferred to the Fischer Family Trust at the end of the project. Details of NFER's code of practice which includes data security can be found here: https://www.nfer.ac.uk/about-nfer/code-of-practice/code-of-practice_home.cfm. No individual school or pupil will be identified in any report arising from the research

Will we be kept informed about the outcome of the study?

Yes. We are committed to disseminating the results of all our research in an open-access framework. All our publications are free to download. We will organise a conference at the end of the project to which all participating schools will be invited. You will hear the results of our research and also have the opportunity to feedback how you feel the intervention worked.

Researchers

Simon Burgess. Professor of Economics at the University of Bristol and Director of CMPO. Simon has worked on education issues for over ten years. He has just finished as Director of the Department for Education's Centre for Understanding Behaviour Change. He has been working with the EEF for three years, running interventions in schools and using RCT research designs.

Shenila Rawal. Shenila is a quantitative researcher specialising in education, teacher labour markets and issues of gender and poverty. She completed her PhD on teacher quality and effectiveness at the Institute of Education, University of London. She has worked on several education and particularly

teacher-related projects for organisations such as the Department for International Development (DFID), the World Bank and Cambridge Education.

Julia Carey. Senior Project Manager at CMPO. Julia has extensive experience of delivering projects into schools, including flagship initial teacher education programmes and RCTs. Julia completed her MSc in Strategic Management at the University of Bristol including a dissertation on school leadership. She is a fellow of the Chartered Management Institute. Day-to-day operational activity will be coordinated by a Project Co-ordinator specifically recruited to support the project, and managed by Julia.

Project partners

The Centre for Market and Public Organisation (CMPO) is an internationally renowned research centre specialising on public service reform. It has carried out path-breaking research in education for over a decade: <http://www.bris.ac.uk/cmipo/>

The Teacher Development Trust (TDT) is a non-profit organisation promoting world-leading approaches to teacher learning: <http://www.teacherdevelopmenttrust.org/>

RANDA Solutions is an INC. 5000 software firm based in Franklin, TN serving the education sector. RANDA seeks to transform education by providing stakeholders timely, accurate and useful education intelligence: <https://randasolutions.com/>

The National Foundation for Educational Research (NFER) is an independent charity working to provide evidence that improves education and learning and as a result the lives of learners: <https://www.nfer.ac.uk/>

The Education Endowment Foundation (EEF) is an independent grant-making charity dedicated to breaking the link between family income and educational achievement, ensuring that children from all backgrounds can fulfil their potential and make the most of their talents: <http://educationendowmentfoundation.org.uk/>

Appendix F: Randomisation SPSS syntax

School- and department-level randomisation

Title 'Randomisation for Bristol teacher observation trial (EFTO)'.

subtitle 'School randomisation 19th September 2014'.

*The SPSS code below sorts cases into random order within strata and allocates the first half to group 1 and the second to 2.

GET DATA

/TYPE=XLS

/FILE='k:\lefto\file for nfer.xls'

/SHEET=name 'Sheet1'

/CELLRANGE=full

/READNAMES=on

/ASSUMEDSTRWIDTH=32767.

*Remove the two lines with no schools.

select if not missing(urn).

*Remove the Welsh school as EEF do not work in Wales.

select if urn ne 401709.

*Populate the missing values as per message from Simon Burgess on 15th Sept 2014.

do if urn=136105.

compute perf_hi=0.

compute fsm_hi=1.

compute white_hi=1.

end if.

*Check for duplicates.

sort cases by urn.

```
match files file=*/first=f/last=l/by urn.
```

```
cross f by l.
```

```
temp.
```

```
select if any(0, f, l).
```

```
list vars=urn perf_hi fsm_hi white_hi.
```

```
**pattern for strata.
```

```
compute pattern=sum(100*perf_hi, 10*fsm_hi, white_hi).
```

```
freq pattern.
```

```
*Stratified randomisation.
```

```
set rng=mt, mtindex=210684.
```

```
compute rand2=rv.uniform(1,2).
```

```
sort cases by pattern rand2.
```

```
compute allocation=$casenum.
```

```
*Allocate the cases.
```

```
recode allocation (1, 2, 3, 8 thru 14, 22 thru 28, 37 thru 42, 50 thru 54, 60 thru 66, 75 thru 81, 89, 90=1)  
into group.
```

```
recode allocation (4, 5, 6, 15 thru 21, 29 thru 35, 43 thru 48, 55 thru 59, 67 thru 73, 82 thru 88, 91,  
92=2) into group.
```

```
*Allocate the odd cases.
```

```
set rng=mt, mtindex=31048.
```

```
compute rand3=rv.uniform(1,2).
```

```
sort cases by rand3.
```

```
temp.
```

```
select if any(allocation, 7, 36, 49, 74).
```

```
list vars=allocation/format=numbered.
```

```
recode allocation (7, 36=1) (49, 74=2) into group.
```


freq group.

cross perf_hi fsm_hi white_hi by group/cells=count col.

ADD VALUE LABELS Group 1 'Intervention' 2 'Control'.

*Randomise dosage of departments in intervention schools.

set rng=mt, mtindex=270452.

compute random=rv.uniform(1,2).

sort cases by group random.

if \$casenum le 23 dose=1.

if \$casenum gt 23 and \$casenum le 46 dose=2.

ADD VALUE LABELS dose 1 'High maths low English' 2 'High English low maths'.

freq dose.

CROSSTABS dose by group.

SAVE OUTFILE='K:\EFTO\EFTO randomisation.sav'.

SAVE TRANSLATE OUTFILE='K:\EFTO\EFTO randomisation.xls'

/TYPE=XLS

/VERSION=8

/MAP

/REPLACE

/FIELDNAMES

/CELLS=LABELS

/DROP=random f l pattern rand2 allocation rand3.

output save outfile='K:\EFTO\EFTO randomisation.spv'.

Teacher-level randomisation

Title 'Randomisation for Bristol teacher observation trial (EFTO)'.

subtitle 'Teacher randomisation 9th October 2014'.

```
GET DATA /TYPE=XLSX
```

```
  /FILE='K:\EFTO\Aggregated data set unencrypted.xlsx'
```

```
  /SHEET=name 'Sheet1'
```

```
  /CELLRANGE=range 'A1:C581'
```

```
  /READNAMES=on
```

```
  /ASSUMEDSTRWIDTH=32767.
```

*Are the TIDs unique?.

sort cases by tid.

match files file=*/first=f/last=l/by tid.

cross f by l.

delete vars f l.

sort cases by urn me tid.

dataset copy teachers.

*Check schools are all from the intervention group.

get file='K:\EFTO\EFTO randomisation.sav'/keep=urn group.

sort cases by urn.

match files file=teachers/table=*/in=inrand/by urn.

freq inrand group.

***Stratified randomisation of teachers.

*If we ensure schools are in random order.

*And within schools departments are in random order.

*And within departments teachers are in random order.

*We can allocate group in sequence.

```
aggregate outfile=*/break=urn/nteachers=n(tid).
```

```
freq nteachers.
```

```
set rng=mt, mtindex=30001.
```

```
compute schrand=rv.uniform(0,1).
```

```
dataset copy schools.
```

```
dataset activate teachers.
```

```
aggregate outfile=*/break=urn me/nteachd=n(tid).
```

```
freq nteachd.
```

```
set rng=mt, mtindex=30002.
```

```
compute deptrand=rv.uniform(0,1).
```

```
dataset copy depts.
```

```
match files file=teachers/table=schools/in=insch/by urn.
```

```
freq insch.
```

```
match files file=*/table=depts/in=indept/by urn me.
```

```
freq indept.
```

```
set rng=mt, mtindex=30003.
```

```
compute teachrand=rv.uniform(0,1).
```

*Randomise.

```
sort cases by schrand deptrand teachrand.
```

```
compute threes=3*trunc(($casenum-1)/3).
```

```
compute group=$casenum-threes.
```

```
list cases=from 1 to 20.
```

```
freq group.
```

cross urn by group/me by group.

recode me ('E'=1) ('M'=2) (convert) into menu.

compute meurn=1000000*menu+urn.

cross meurn by group.

add value labels group 1 'observer only' 2 'observee only' 3 'observer and observee'.

sort cases by urn me tid.

save outfile='k:\efto\EFTO teacher randomisation.sav'/keep=urn me tid group.

SAVE TRANSLATE OUTFILE='K:\EFTO\EFTO teacher randomisation.xlsx'

/TYPE=XLS

/VERSION=12

/MAP

/REPLACE

/FIELDNAMES

/CELLS=LABELS

/DROP=nteachs schrand insch nteachd deprand indept teachrand threes menu meurn.

dataset close all.

Appendix G: Statistical Analysis Plan

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	23/12/16

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 Key Stage 2 to maths GCSE accounting for student gender, major ethnic group and FSM), and school English VA (English Key Stage 2 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 2014.¹⁶ Specifically, and in terms of the September 2015 edition of NPD Data Tables¹⁷, 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.

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. Key Stage 2 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 model¹⁸. 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.

¹⁶ 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.

¹⁷ <https://www.gov.uk/government/publications/national-student-database-user-guide-and-supporting-information>

¹⁸ In the protocol, Key Stage 3 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.

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 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 Key Stage 2 score¹⁹ 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 study²⁰. 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

¹⁹ The 2010 Key Stage 2 boycott may affect the 2015 Year 11 analysis. If so, a measure that incorporates teacher assessment may be needed. (Annotation: KS4_VAP2TAENG_PTQ_EE and KS4_VAP2TAMAT_PTQ_EE were used.)

²⁰ And over two years in the case of the year 10 cohort that starts the trial in October 2014.

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 Key Stage 4 points score in maths with Key Stage 2 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

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 Key Stage 2 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 (Key Stage 2) 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.

You may re-use this document/publication (not including logos) free of charge in any format or medium, under the terms of the Open Government Licence v3.0.

OGI This information is licensed under the Open Government Licence v3.0. To view this licence, visit <http://www.nationalarchives.gov.uk/doc/open-government-licence/>

Where we have identified any third party copyright information you will need to obtain permission from the copyright holders concerned. The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Education.

This document is available for download at www.educationendowmentfoundation.org.uk



Education
Endowment
Foundation

The Education Endowment Foundation
9th Floor, Millbank Tower
21–24 Millbank
London
SW1P 4QP
www.educationendowmentfoundation.org.uk