The Wellcome Trust Case Study of the Implementation of a Randomised Controlled Trial (RCT) in Education

Exploring Perceptions of the Primary Science Specialist (PSS) Continuing Professional Development (CPD) Evaluation

wellcome trust

THE UNIVERSITY of York

Department of Education

The Wellcome Trust Case Study of the Implementation of a Randomised Controlled Trial (RCT) in Education

Exploring Perceptions of the Primary Science Specialist (PSS) Continuing Professional Development (CPD) Evaluation

Final Report

November 2014

Case Study Project Team:

Jeremy Airey

Judith Bennett

Eleanor Brown

Contents

Executive Summary

Conducting RCTs

Recruitment

The key issues for recruitment were the exacting criteria teachers had to meet to qualify for participation in the intervention and its evaluation, and the time frame in which the recruitment took place. In addition, there was an issue with getting schools to accept the randomisation of group allocation, being able both to commit to 24 days of CPD and at the same time be content to be involved for no immediate benefit.

Statistics and Sample

A number of factors may have influenced the nature of the sample and hence the findings of the RCT. These can be termed Hawthorne-type effects. There was evidence from many schools that being involved in the research had raised the profile of science, and that being interviewed and doing subject knowledge assessments had influenced the teachers to think more about science and possibly look for CPD elsewhere, particularly if their school had been allocated to the control group. Indeed, the nature of the sample itself tended to be biased towards schools which were already prioritising science in some way.

Measuring Impact

Capturing and quantifying the outcomes of the CPD was a major challenge. Using an instrument comprising items from an existing bank (Key Stage 3 Standard Assessment Tests, SATs) enabled a broad measure of subject knowledge to be obtained. However, as the CPD and the evaluation evolved, the desirability of a more finely-tune instrument became apparent. For this to be possible, more communication between the CPD providers and the evaluators is needed to ensure a better match between the focus of the instrument and the specific intended learning outcomes of the CPD.

Participating in RCTs

Reflecting Teacher Knowledge through Assessment

There were mixed responses from the teachers about whether they found the assessment was a good reflection of their subject knowledge. Some found it useful, while for others the assessment was seen as too difficult, and in some cases disempowering. Indeed, many thought it was a not a good reflection of their knowledge or ability to teach it and those in the intervention groups often said that the assessment did not allow them to demonstrate what they had learned on the course. There were some issues with the implementation of the assessment, with questions around the fidelity with which the assessments were conducted.

Reflecting Pupil Knowledge through Assessment

With the pupil assessments, the main issues raised by the teachers were that the topics covered by the assessment had often not yet been covered in the class and so their pupils were not able to demonstrate their understanding. Moreover, the use of an exam-style assessment was off-putting for some of the pupils, many of whom had never done exams before and found this way of presenting their knowledge difficult.

Experience and Expectations

Many teachers from all three groups felt they had had a positive experience of being involved in the RCT. However, from some of the control group there were comments about the difficulty of having to put in effort for no immediate reward. Many said that they needed to have the expectations of what was involved more clearly explained to them at the beginning, and some felt that the amount of time commitment involved had been underplayed at the recruitment stage. The importance of good communication, particularly with the control group was noted by a number of teachers.

Key Themes

Challenges of recruiting and the nature of the sample - The barriers to recruitment such as the nature and extent of the criteria the teachers had to fulfil and the time available had considerable implications. The demands on the teachers during the research, to act both as participant and researcher also exacerbated attrition.

Combatting attrition - It is important to militate against attrition, and one way is to ensure good communication with the participants, especially at the start.

The scope of what an RCT can detect - The broad focus on the RCT to improve science knowledge in a wide range of areas made it difficult to measure the outcome.

Communication and collaborative definition of outcomes - Having a qualitative dimension to an evaluation is important as it can be used to inform the design of the RCT and refine the CPD in advance of collecting the quantitative data.

People as participants - The influence of the research on the control group and Hawthorne-type effects are inevitable when doing experiments with human participants. The sample is also likely to be skewed towards those who would be willing to be in the full intervention group, and where that is very intensive, as in this case, that means they are schools that already prioritise science.

Fidelity of data – Teachers approached the assessments in different ways. In some cases teachers did not complete both sets of assessments, or they gave the second round of assessment to a different class.

Recommendations

Specificity - Keep the design well-defined and narrowly focused.

Simplicity – Avoid requiring teachers to be both participants and researchers and asking them to make demands on colleagues as these increase the risk of missing data. Running the intervention and evaluation in parallel reduces the flexibility of the evaluation to respond to changes in the intervention.

Scoping – A pilot phase which gathers some preliminary qualitative data permits the design of a more focused RCT.

Sample – Ensure that the criteria for participation are appropriately selective without being unduly restrictive. Ensure that the commitments being made, and a good understanding of the research project and its design, are signposted clearly to potential participants during the recruiting phase.

1. Introduction

In this report we give an overview of a small case study regarding the use of randomised control trials (RCT) in education. We discuss the perspectives of key informants and draw some conclusions about conducting RCTs with a few tentative recommendations based on the experience of this project.

2. Background to the Study

The use of randomised control trials (RCT) in education is relatively new, and while they are becoming increasing popular, there is still a lot of debate about the use of this method for researching education. An RCT looks at a sample of a population, collecting baseline data, and randomly assigning the sample into groups where some receive an intervention of some kind and others act as a control group which does not receive the intervention.

The Wellcome Trust is currently funding a two-year research project to evaluate the impact of a continuing professional development (CPD) programme for primary school science co-ordinators and their colleagues. This is a large scale intervention aiming to improve science knowledge and pedagogy amongst primary science coordinators with no more than GCSE science. The Primary Science Specialists (PSS) project CPD is provided by the National Science Learning Centre (NSLC).

Part of the evaluation project involves undertaking an RCT of the impact of the CPD. This aims to test the effectiveness of the intervention through pre and post testing of three groups of schools: Full intervention schools receiving 24 days of CPD; Partial intervention schools receiving 4 days of CPD; Control schools, which receive no CPD provided by the NSCL over the two year period. The evaluation aims to determine the success of the CPD based on the measurement of improvements in the science knowledge of the coordinators. In addition the RCT measures pupil achievement and attitudes to science and teacher confidence in completing science assessments. A teacher colleague in each school is also tested as a measure of science leadership spreading through the school and the impact of science coordination. There is also a qualitative strand with one third of the schools, selected equally from across each of the three groups. Interviews with the science coordinator, a teacher colleague and a senior leader, as well as a lesson observation and a focus group with pupils, provide an additional dimension to the data.

In recent years there has been a growing interest in RCTs as an evaluative tool, and the education research community has been encouraged to adopt such an approach to make educational research 'more rigorous'. Despite this, there is little evidence to suggest large-scale uptake of RCTs in educational research. This may be due to a number of reasons; including (a) resistance from certain sectors of the research community, for epistemological or methodological reasons or because of a sense that depriving some pupils of the opportunity to participate in a new educational initiative is unethical (b) a lack of familiarity with the approach in practice, both from the researchers and in order to recruit participants and (c) the practicalities of implementing an RCT and the difficulties of measuring an intervention in a way that accurately demonstrates the impact. Indeed, there are practical, ethical and methodological considerations to using RCTs in education, and there is a range of stakeholders whose views have a potential impact on the utility of this type of evaluation.

The Wellcome Trust is keen to ensure rigorous evaluations of interventions in science education, and this is the first RCT in education it has funded. Therefore, it is seen as an experiment in itself and a useful learning experience. For this reason, the Department of Education at the University of York is currently being funded by the Wellcome Trust to conduct a small case study investigation of the PSS CPD evaluation, in order to draw out key messages about the use of RCTs in educational research. The evaluation of the CPD programme for primary school science co-ordinators offers a very useful opportunity to gather information on the process of undertaking RCTs. Such information is likely to be of interest to funders of research, policy-makers (particularly those advocating the use of RCTs), educational researchers and those working in schools.

3. Objectives and Design

The main objective of the project is to explore the perceptions, based on experience, of a range of stakeholders about the use of RCTs in educational research. We aim to consider the aspirations for the PSS evaluation from the perspective of funders, providers and evaluators; the challenges of conducting an RCT; the views of teachers and providers of being involved in an RCT; the extent to which RCTs as a method are understood in schools; the implications for schools of randomisation and their experience of being selected for the three different groups. The intention of this short report is to provide guidance for funders, policy-makers, educational researchers and school-based staff on issues to be considered when setting up, conducting and participating in an RCT.

The project has a case study design and draws on qualitative data, through in-depth interviews and focus groups with key stakeholders. Therefore, the project began with a review of the literature about using RCTs in education and some consideration of other research that has used this methodology. This informed the design of the research instruments. In essence, the interviews seek to establish views of RCTs in general, and reflections on the specific RCT being undertaken, including views on the effects of being selected, or not selected, to experience the intervention and the experience of having their confidence and subject knowledge measured through the RCT.

Interviews with evaluators, funders, recruiters and providers were designed to gather their opinions and perspectives on the use of RCTs. The participants referred to as evaluators were those involved in collecting the RCT data, funders included a range of staff from the Wellcome Trust, the recruiters were people involved in the recruitment of teachers to the project, both from the University of York and from the National Science Learning Centre (NSLC), the providers were staff involved in the CPD course based at the NSCL. Conducting a focus group with providers aimed to generate discussion about experiences of being involved in an RCT evaluation, exploring practical and methodological issues. Interviews with teachers were already being conducted by the evaluation team as part of the qualitative element of the evaluation. Therefore, one question was added to the interview schedules on the use of RCTs in both rounds of interviews, in order to generate data about the teachers' experiences. This generated data from a sample of teachers involved in the PSS evaluation with schools in each of the three groups. Finally, an interview schedule was designed for use during phone interviews with teachers who had withdrawn from the project.

This case study uses in-depth interviews to generate qualitative data. The interviews were all transcribed and coded using NVivo10. These codes were then used to identify themes which provided a structure for the report.

3 Conducting RCTs

In this section we report on the experiences and messages emerging from the perspectives of those involved in conducting the RCT. That includes the voices of the funders, the providers, the recruiters and the evaluators involved in the process to different degrees. The perspectives of the teachers participating in the RCT are discussed in section 4 below. Here we look at the RCT in terms of the experiences of recruiting and retaining teachers, managing the statistical analysis and maintaining the sample size, and the ways that learning was measured through the RCT.

From the perspectives of the funders, providers, recruiters and evaluators involved in this project it is commonly recognised that there is a lack of good evidence about what works in education, and RCTs are seen as a way to overcome this. Having robust evidence offered by an RCT enables governments and funders to make informed decisions and policy choices. They are a sophisticated tool that uses statistical analysis which, if it shows there is a significant difference caused by an intervention, can offer very strong evidence. A strong quantitative measure is considered "more effective to influence policy" (Mary - Funder). However, due to the nature of the questions it is able to answer, there is of course a chance that it may tell us that something does not work, and when this happens RCTs are unable to offer an explanation about in what ways, to what extent, how or why something did or did not work. It is therefore seen by some to be a "kind of a gamble" (Rob - Evaluator), and it is an expensive gamble to take.

This project was to some extent seen as an experiment in conducting an RCT in education, and therefore as a learning process about the way to conduct an RCT in this context. It was recognised that as an experiment it might not work, but that as a funding body, the Wellcome Trust would "learn by giving it a try" (Felicity - Funder).

3.1 Recruitment and Retention

There were some significant difficulties with the recruitment and retention of teachers to the evaluation. These were anticipated to some extent and in many ways overcome. However, there are some messages about ways these might be dealt with in future projects.

3.1.1 Randomisation

A key difficulty for schools was accepting the randomisation of group allocation. This is clearly impossible to avoid with an RCT, but it is worth communicating to schools an understanding of the importance of randomisation, as there was evidence that this posed a significant barrier to involvement. Many would say they would be happy to be in the full or partial intervention groups but not the control, while others did not want to commit to the full 24 days of CPD required by the full intervention. Teachers found it hard to understand the importance of randomisation and did not like having to accept the gamble.

> *They didn't want to be in something where they didn't know which group they were going to be in, that was the bottom line, that was it for many schools. (Sally - Recruiter)*

There is an argument, then, for more communication with schools about the benefits of research, particularly on the scale of an RCT, and even for requirements to participate:

> *I think at the moment the emphasis for RCTs is too much focussed on the evaluator trying to find schools, rather than schools seeing the benefit to them. … I would like to see engagement with educational research as being part of a requirement for a good or maybe… well, certainly for an excellent but maybe even for a good school, and that way, as I say, maybe if you said a good school, or an excellent school, needs to take part in educational RCT, maybe just educational research, once in 5 or 6 years, the pool available then would become much larger. (Laurence - Evaluator)*

The time pressure added to the difficulties of recruitment, with a lot of work happening over the summer holidays and some noted that there is often a lot of movement in primary schools between July and September, meaning that sometimes teachers were not then teaching a year group that would enable them to participate in the study (Christina – Provider). Indeed, another issue mentioned by recruiters was the amount of criteria that schools had to comply with in order to participate. This meant that even when schools were keen there were still reasons that they were not able to take part, and this added pressure to recruitment.

> *This was what was difficult in a way, you'd get them quite excited by this prospect and then you'd say, "Well have you got qualifications above GCSE?" "Yes." "Oh I'm sorry you can't do it." ... So it really was the criteria narrowed the field massively. (Sally - Recruiter)*

3.1.2 Sample Size

Given some of the difficulties of recruitment, questions arose about the size of the sample:

> *Now originally, we wanted to use a larger sample size. For various reasons relating primarily to recruitment, the sample size was whittled down from, I think, originally about 127, something like that, down to about 90, and… that has proven, I think, an issue that needs to be considered for future RCTs. (Laurence - Evaluator)*

Perhaps the number of groups in the randomisation exacerbated this problem, with additional pressures on the sample size with the need for three groups. This led to worries regarding the power of the calculations, once the original sample size was smaller and attrition was taken in to account. This had repercussions for the statistical analysis and the possibility of missing differences between the groups.

> *With three groups, of course recruitment is even more difficult than with two groups. But we've still managed to maintain the statistical power, we have a formula in there, which I think, I can't remember whether it was 0.6 or 0.8, but either of those was acceptable for an RCT. And, even when we had some attrition of schools dropping out, we did careful calculations and we were able to reassure Wellcome that we still had that statistical power. … In terms of what it really means, is … would it be possible then for someone to argue that, if the RCT is showing that there's no statistically significant difference, it could be that the difference lies in this 40% of things that haven't been measured. (Rob - Evaluator)*

Finally, there are also issues in terms of the number of variables that can be accounted for, and how the sample size impacts on these.

> *There are differences in local authority aspects as well, and now we've got such a range of types of school that makes a huge impact as well. (Sarah - Funder)*

3.1.3 Demands on Teachers

Another issue with retention was that there were high demands on the teachers, not only in terms of their time, but also because they had to sit subject knowledge assessments. This could feel threatening to a teacher, especially the teacher colleagues; such a method is not often used, so may have been unfamiliar to the teachers. It is quite a sensitive thing to do and could be disempowering to the teachers as we discuss in section 4.2.1.

3.2 Managing Statistics

With a sample size smaller than anticipated and some issues regarding the nature of the control group, managing the statistics was not an easy task. The evaluation team were confident that the power of the statistics remained strong and that variables were controlled. However, there were some things that emerged from the interview data that highlight areas which are difficult to control for, such as the nature of the sample and Hawthorne-type effects. We discuss these here from the perspectives of those conducting the study and in section 4 from the perspective of the teachers.

3.2.1 Sample Skew

Given that the teachers at the point of recruitment had to sign up, in principle, for 24 days of science CDP, they were inevitably teachers or schools with an interest or commitment to science.

> *I mean you're only going to sign up to this, really, with the chance you might be asked to do 24 days of science CPD across a year, if you do think science is really important and valuable. (Rebecca - Evaluator)*

So many of the schools were already enthusiastic about science and were hoping they would be in the intervention group. Therefore if they were in the control group they would often look for other CPD opportunities for science leaders.

> *There could be some schools, and I went to a school where they were in the non-intervention group, and the teacher was really really enthusiastic about science, was putting herself through all kinds of training programmes, was doing a Masters in science… everything she, you know, she wanted to improve. Now you could say "well, that has had just as big an impact as someone who goes on 20 days of CPD". (Laurence - Evaluator)*

This enthusiasm made distinguishing between groups A and B even more challenging, since given that they were getting some CPD, their approach and response to it was often perceived to be determined more by the commitment of the individual teachers than the effect of the course.

> *The teachers that you've got on here… the teachers who all signed up to this, want to make a difference to science education in their schools. And they're enthusiastic about it, so if you're looking at the difference between group A and group B, in terms of their*

commitment to making a difference, there wasn't one at all, because they'd already signed up to it. So you've got a variable there, the teachers' own motivation, and willingness and … And the group Bs that I worked with, did that, they… they were equally as engaged and committed to make a difference in their school as the group A teachers were. (Focus group - Providers)

3.2.2 Hawthorne-type Effects

The simple fact of being involved in a study about science inevitably raised the profile of science in all the schools.

I suppose once head teachers or senior leaders in school get an idea that there's a focus on an area, then they're going to possibly start focussing on that stuff, to develop those areas, because they think "well, actually, somebody else thinks it's important, so therefore we will do something about it" (Focus group - Providers)

Particularly with the control group, it could be seen that since they were doing a lot of evaluation work and participating in the project, the focus on science could have made them more engaged with science and therefore impacted on their results.

There's so many variables that you can't account for, and the fact that you're taking a control trial of people you're telling them that they're part of a control is possibly going to have some effect on their teaching of that subject. I think it's very hard to… work around that. (Focus group – Providers)

3.3 Measuring Impact

Perhaps the most important issue, however, when conducting an RCT is the selection of outcome measures, a discussion of the ways impact, in this case learning, is measured, and what aspects of that learning need to be captured. There was discussion of the extent to which the intervention should focus on science knowledge or pedagogical content knowledge (PCK), or even approach to science in school. This led to difficulties in capturing the outcomes. We found a number of themes emerging in terms of the experience of capturing learning. These included the way that learning could be quantified and how the qualitative aspects of learning could be rationalised alongside the RCT.

3.3.1 Capturing and Quantifying Learning

One of the questions that that the funders wanted to answer at the outset was about the balance between pedagogy and subject knowledge, wondering whether CPD should offer "more pedagogy and less subject knowledge, or vice versa" (David - Funder). It is difficult to see how the RCT would be able to answer that question, since increased subject knowledge was the measure, and therefore focused on in the course, but the RCT could not show whether a focus on pedagogy would have been better or worse against this measure.

Given the specificity required to capture the learning through an RCT, it may have been useful to develop bespoke outcome measures. However, it was recognised that with the time constraints this would never have been feasible for this project.

> *… within all those constraints I think we got as good a measures as we could. If we had more time, to develop, or were able to bring in more people, then I think we would have been able to tailor those assessments which could have ended up with richer information. (Mary - Funder)*

Certainly, pre-testing the instruments more would be seen as a benefit for future studies (Sarah – Funder). The decision to use SATs assessments made a lot of sense given the time constraints, but it was also recognised that both the content and style of the assessment may not have enabled all of the learning to show through and that there were aspects of Key Stage 3, which may come up in the assessment, that do not really impact on Key Stage 2. The very nature of using an examination style paper was seen as a drawback.

> *… they are about regurgitating knowledge, actually what we're doing is giving a very rich conceptual understanding, which isn't properly accessed by the SATs assessments. (Mary - Funder)*

There was a lot of discussion about the extent to which the providers and the evaluators communicated about the objectives of the learning and the content of the assessments. The design was such that the providers could not know what the assessments would measure and the evaluators kept a distance from the course, so that they did not know what was being covered. However, there may be an argument for having clearer communication about the objectives and focus of the course, as 24 days is not enough to cover an entire Key Stage 2 syllabus, let alone also Key Stage 3, so it would always be difficult for the teachers to show improvement when the measures were so broad.

Nevertheless, attempting to define a quantifiable measure that would have been more appropriate, even with the benefit of hindsight, is extremely difficult. Whatever measure one uses, there will always be things that are missed by an RCT, precisely because it measures one particular outcome. The providers had a number of ideas that they suggested might be able to detect the learning they had witnessed as a result of the course, such as true false tests to measure misconceptions or conducting structured observations, but they recognised that often these were things that were difficult to quantify.

> *He said "so I know I'll have got that part of the testing wrong, but that's not a reflection on the course, because I know we did it on the course, and I know if I'd been teaching that in my class, I would have known it absolutely fine, because I would have known I was teaching Light that day, so I would have prepared for it, whereas, when you're given a cold exam, you don't know what's going to be on there" So, unless you've gone through every single thing, and re-revised it… And a lot of our teachers also just said "come into school, come and see us teaching, come and talk to the children, come and talk to our heads" But obviously that's difficult because it's not… it's emotional, it's subjective, it's not… you can't quantify what they've learnt. (Sandra - Provider)*

3.3.2 Qualitative Aspects

Indeed, capturing more qualitative aspects of learning became a common feature of debate. There was a great deal of qualitative evidence as well as anecdotal comments that the course had benefited teachers immensely, but it was difficult to capture some of this detail quantitatively.

> *Thinking about the learning, it is interesting, those teachers knowing they're special, they know they've been on this course, they tell us that it's a life changing experience for them. It is interesting how hard it is to capture that numerically. (Mary - Funder)*

This led to a general consensus that with RCT research there was a need for a mixed methods approach. Indeed, the qualitative data allowed the research to answer different types of questions and add detail to the study. It was agreed that ideally good research needs "a bit of both" (Sarah - Funder).

> *What it illustrates, is what one knew already, really, I suppose, that an RCT on its own is not enough. Because there are collateral things happening around this intervention which mean that I'm jolly glad that we've got some qualitative work running alongside the RCTs. … so you have to surround… you have to design your RCT but then surround it with a whole bunch of intelligence, you know, it's like having spies going out there, and rooting out the undergrowth as well as having the main assault going on there in the theatre of action. … And I certainly wouldn't want to pin all our hopes for future educational decision making on the outcomes of RCTs alone. Persuasive though they can be. (David - Funder)*

However, in some ways the qualitative and quantitative data seemed to contradict each other. There was a lot of data about ways in which the course was perceived to have improved teacher practice and the attitudes to science throughout the school had changed as a result of the intervention.

> *She told me who it was … and the teacher had been on this course here, and so… less than 12 months on, this teacher is making a huge impact across the school. So much so that her colleagues are now coming out on courses, and talking about the stuff that is going on in the school, and… everybody's impacted, and how engaged the children are. (Focus group – Providers)*

However, there is an argument that teachers often self-report improvement after CPD, while in fact their knowledge had not improved as much as they thought, so some claim that this is not as reliable as a quantitative measure.

> *What teachers also tend to self-reflect and report on, is a change in their practice. And I think what we're seeing, is that there hasn't been as much change as they might think, and in terms of improvements in subject knowledge, again, frequently referred to in the literature as resulting from CPD, don't seem to be borne out by the hard evidence. (Laurence - Evaluator)*

> *Well, I think that's why you have to have the randomised element because very often peoples' impressions of how well an intervention is working are not reliable, are not accurate. (Luke - Funder)*

On the other hand, teachers can be discerning about CPD, and often report when they have not learnt anything, or that is has not improved their practice. Indeed, the benefit of talking to teachers about their perceptions is that often you discover unexpected outcomes, which the RCT was not set up to detect.

> *You could conclude that it hasn't worked, you could conclude that it hasn't worked in the thing that you thought it would do, but it's done something else. So for example you might find that actually pupils' achievements aren't that much greater, but my goodness, they're much more interested in science. And I think that probably is something that's coming out of this. That's a perfectly valid finding. (David - Funder)*

4 Participating in RCTs

In this section we look at responses from teachers about being involved in an RCT, how they felt the assessment reflected their knowledge and confidence and the knowledge and attitudes of their pupils. We discuss their experiences of being involved in the RCT. This is divided into five sections: teachers' perceptions of RCTs in general, their experience of completing the assessments themselves, their perceptions of the pupil assessments, their experiences of being involved in the evaluation in general, and some issues arising that may have influenced the findings.

4.1 Perceptions of RCTs

The teachers were generally positive about the idea of RCTs as a way of conducting research. They had generally not given it a great deal of thought and often had a superficial understanding; however, they could usually see the benefits. While some of them had found the logistics of having three groups and having to be selected randomly difficult, they generally understood the need for this.

> *Yeah well randomised controlled tests are the best way of doing things and people don't like doing them in education because nobody wants to be the group that's left out … and it's getting your head round as a school leader the fact that children aren't disadvantaged … because they're getting what they would have been getting anyway, but you are trying to make it a bit more evidence based and a bit less gut feeling based and I'm very happy with that. (79 Full SM)*

Others were more sceptical, and felt that it was difficult to rely on this data given the level of subjectivity involved in having participants with different characteristics and levels of commitment.

4.2 Perceptions of Teacher Assessment

A key aspect of the teacher interviews examined the extent to which they felt the assessment captured their knowledge and confidence. There were mixed responses in both cases in terms of how well they felt the assessments reflected their feelings of efficacy. These are explored below.

4.2.1 Science Knowledge

The teachers were assessed in terms of their subject knowledge and their confidence in answering the questions. It is interesting that many of those who said it was a good reflection of their subject knowledge were in the control group.

> *Yes, I think a reasonable one, I mean I felt doing them, it was quite a while since I did them, but I remember when I did them thinking they were probably … most of the questions were about GCSE level, so obviously quite a bit beyond what we would teach in a primary school, but obviously probably around the level of knowledge you needed to be able to teach more basic knowledge, if that makes sense? You always need to know quite a bit more than what you are teaching, don't you? So yes I think it was reasonable, yes. (18 Control TC)*

The full intervention group tended to feel it was a reasonable reflection at the beginning, but that it did not necessarily reflect their improvement, commenting that the assessment was unable to capture the learning that they had taken from the course.

> *I think they were, I think they were fair tests at the beginning, they tested our subject knowledge fairly at the beginning, of where we were starting out from. (58 Full SC)*

> *And I know the, the viewpoint from everybody on the course, quite a few people commented about the exam papers and couldn't see the connection to how to, how does that measure what that course has done for us. (86 Full SC)*

Difficult

Many of the teachers were struck by how difficult they found the assessments and that it was generally considered to be GCSE level. It is interesting to note that the assessments were Key Stage 3, and therefore the teachers perceived them to be a higher level than they were, which was indeed only the next step up for their pupils. This may be an indication that the teachers underestimate what their pupils need to be achieving by the end of primary school.

> *Oh, terrible. It was awful. … I mean there was a lot of stuff that I thought I did in GCSE which, obviously, was a long time ago. So, I kind of recognised stuff but didn't have the knowledge to be honest with you, a lot of the answers I did just guess. (38 Control SC)*

Indeed, many raised the point that they were difficult because they addressed topics that they never cover in primary school and therefore had generally not thought about since they were at school.

> *I think a lot of it was quite a lot beyond what obviously we teach at primary and showed us that, you know, it's been a long time since we did a lot of that stuff at high school. (87 Control SC)*

Useful

Despite finding them difficult many teachers commented that the assessments were useful and they had found it interesting to discover the gaps in their knowledge and highlighted areas for development.

> *I think it was quite interesting because you kind of think, yes I know a fair bit about Science, but actually it did – those particular questions did make you think. Well, could I teach that? Do I actually know the definition of what that is? So it did make me question some of the things that you think you've got a handle on, but actually perhaps I would want to check it up before I delivered it. So I thought it was quite useful to hone in on your own knowledge really. (79 Full SC)*

Some teachers in the full intervention group reported feeling that they had done better on the second assessment, and felt more confident in general with using scientific language.

> *Well the first one was shocking. Before the CPD it was shocking. I think I did much better on the second one. (53 Full SC)*

Disempowering

On the other hand for some teachers the experience was disempowering and made them nervous and less equipped to teach science. Some said it made them feel worried or as if they were not good enough. One teacher said that it was detrimental because it may her feel that she did not have the appropriate knowledge, even though she saw it as above the level she had to teach, another found it demoralising.

> *I remember that the teacher test paper was hard. I had the feeling that there were some worrying gaps in my science knowledge. This was a bit demoralising. Also, I have to say that I have never used that knowledge or content in lessons. (43 Partial SC)*

One senior leader also saw the assessment as potentially daunting for staff.

> *So I did feel "Oh, you put this in front of some staff" and not if you teased it out of them they wouldn't necessarily know a lot of it but I think the format and sat down and doing it like that would've been very daunting for some of the staff. (94 Full SM)*

Not a good reflection

Many teachers felt that it was not a good reflection of their knowledge and in the case of the two intervention groups, that the assessments did not necessarily reflected their learning. The fact that the assessment was pitched at secondary level was unhelpful for many.

> *I don't think I would have seen the point in it really, you know, what's testing the dregs of my GCSE knowledge … would that have helped the children learning. I do understand that, you know, you have to have an understanding in more depth in order to be able to teach, but I think you do that anyway. (79 Full TC)*

Many teachers felt that what came up on the assessment was not necessarily relevant for what they needed to know to teach, nor what they had learned on the course.

> *Well, I think my Science subject knowledge, they were probably a good reflection of that but not in teaching because I don't teach a lot of what they were asking. There was a very heavy Physics based one, if I remember, and I don't tend to do a lot of that. The subject matter wasn't quite relevant really, I suppose. (68 Full TC)*

Indeed, many teachers commented on the fact that they always prepare for lessons and 'brush up' on areas they are not sure about. This was not taken into account by the assessments.

> *I tend to find that if new things come up in the subject that we need to then be teaching, we will then go away and actually research it, so we are prepared ready for the lesson. So we might not necessarily have that information already inside us, but we will make sure that we find out what it is, so that we can teach the children it to the best standard. (22 Control SC)*

For others the very nature of testing their understanding with an 'exam-style' assessment went against their understanding of good education. Learning and understanding is about more than knowing facts.

> *But I think as far as primary school, it's more about how you implement the learning and give the children opportunities to find out rather than filling them full of knowledge. (79 Full TC)*

> *I felt more confidence, you know, talking to people about science and doing experiments and explaining things and explaining why they happen, but then putting it on paper I found hard but personally as well for me, I don't like exams. It's not my kind of style of learning. (86 Full SC)*

4.2.2 Teacher Confidence

The confidence part of the assessment was less of a concern for the teachers. Most felt that the confidence rating had accurately reflected how they felt about answering the questions, although not with that entire subject area. They were clear to point out that that did not necessarily reflect their confidence or knowledge of a particular topic, but rather of that specific question.

> *So the confidence was … I interpreted the confidence relating to those questions, not my subject knowledge in that particular topic. (18 Control TC)*

On the other hand, some teachers did not feel the confidence rating was a useful measure because they felt that it represented only facts, rather than an understanding of the subject and an ability to teach it.

4.2.3 Implementation

From questions around the implementation of the assessment a few key issues emerged. These were issues of consistency regarding how the assessment was perceived and administered and the time given to the doing the assessments. In that sense there is a question around the fidelity of the completion of assessments so that they measure what they were intended to. In terms of consistency, some teachers did the assessments watching the television, others at school, and in many cases there was also a difference between doing the first and second assessment.

> *Hmm, I think I value the fact that we had to be tested. I'm not quite sure whether everybody did their tests in the same way as some of us. So, you know, sort of, sitting at home in front of the TV filling in bits and pieces while they're on the laptop, to make it look good. I don't know whether that was the same as – you've got 40 minutes – do*

> *it in assembly time. Do you know what I mean? So, I'm not quite sure whether – if they're not tested centrally... (3 Full SC)*

There was a lack of clarity about whether it was appropriate to revise for the assessment and whether they should look at their notes from the course in advance. This was particularly an issue for teacher colleagues, where perhaps there had not been time for science leaders to pass on any of their learning or embed practice in the school.

> *I was saying, am I supposed to revise or brush up on it? If I thought, but I haven't done anything in that time, so if I didn't know it then, chances are I wouldn't know it again … so doing that for me with me not really knowing much about what I'm expected to do, I don't think, I don't see the relevance of that with me actually, because I haven't been on any of the courses … I'm not quite sure what I'm supposed to get out of that or what you're supposed to get out of that. (86 Full TC)*

The amount of time dedicated to the assessments also varied significantly. Teachers in the full intervention group were those who most commented on not really having sufficient time to spend on the assessments.

> *… probably the worst headache as a teacher is having to sit the tests … Because, partly timing, it's partly finding, because you have to get cover for that length of time to be able to sit, I know this sounds ridiculous but we just don't have that time. (94 Full SC)*

> *I wasn't expecting maybe the time it took, my own. It took forever. (53 Full SC)*

Teachers were also asked if they thought another measure might be a better reflection of their knowledge. Most referred to the idea that someone should watch them teach, and see the difference before and after the CPD.

> *… I do feel a lot more confident in teaching science. I feel a lot more confident in advising other teachers about science, so I think as a measure, if you'd have seen me before. Just as a normal standard science lesson, and then watched me maybe in the middle of the course, and as you're doing today, watching another one, I think you would see a difference and I would think it would be quite clear in terms of the confidence level of my subject knowledge as well within those topics. (86 Full SC)*

4.3 Perceptions of Pupil Assessment

Here we address how the pupil assessments were perceived by the teachers. They talked about subject knowledge and attitudes. They raised some important issues about timing, in terms of content covered in class, the style of the assessment and language used, pupils not being prepared and problems of combined year groups. There were also discussions of differences in the ways the assessments were administered.

4.3.1 Subject Knowledge

Some teachers were quite accepting of the assessment and felt it gave a reasonable impression of the pupils' knowledge.

The pupil tests were quite efficient. They seemed to have similar questions. They were actually very helpful for our pupils. They were clearly better at some sections of work and this is probably linked to the units of work they had been doing recently. Some pupils are not very good at retaining their knowledge. We would be interested to see the test results, to see the extent of progress in our children. (Partial SC)

Topics covered

There were many comments about whether or not pupils had covered the topics that came up on the assessment and how recently.

With the children's one, there was quite a few of the questions we hadn't taught them yet. They couldn't answer them very well even though some of them could have a guess at them … there were a few which as we go through I'm thinking, "They're not going to know that yet because it's not come up". (Control SC)

This caused an additional problem in combined year groups and mixed classes, where some of the pupils had not covered many of the topics. This led some teachers to ask whether the assessment could really pick up how well they pass on subject knowledge or simply whether or not that topic had been covered with those pupils.

It was hard to judge because a lot of it I felt was – it was like end of year questions and when we did the tests they were at the beginning of the year so there were lots of gaps, but that's because we hadn't covered stuff. So in terms of actual subject knowledge there were things that they wouldn't have known. (Full SC)

Similarly, there were comments about curriculum and the need to ensure the content of the assessments fitted with what the children were learning in class.

Sitting assessments

Moreover, according to some teachers, since most children were not used to sitting this type of assessment in science, and perhaps not in other subjects either, the style of doing what they perceived as an examination did not really give a true reflection of what they had understood in class.

I think probably … the presentation of some of the questions will throw some of the children who haven't been through the SATs preparation. (Full SC)

The children felt that the tests were quite hard – this may be because they are not that used to formalised tests. (89 Control SC)

Many teachers commented that their pupils had found it difficult and really struggled to complete it in the time available. Some said that the assessments did not reflect the way they teach science.

I think the tests are quite difficult because it is subject knowledge, isn't it, and it is very much question and answer, and we don't teach Science that way because we go down very much the content cartoon route and they start with a problem and they off and they find their own answers. So actually having to then put that into a fact style paper

> *was quite tricky for us. You might not have actually seen their full level of understanding. (Control SC)*

> *… the emphasis is on the children doing and the children understanding and not sitting tests they get a piece of paper put in front of them with SATS-type questions and they've just never done anything like that, so it doesn't necessarily reflect. (87 Control SM)*

Indeed, the point was made that doing a science 'exam-style' assessment and 'doing science' are two very different skills. Many children had never done a science exam, which is a skill in itself, and they were assessed on things that had never been taught in class.

> *No. And if that was part of the regular regime they would be prepared for that as part of their learning experience but for us the way we're taking measurements of where they are in science we don't have to show them in that format for them just to learn that skill, because that's a different skill to actually doing science. (87 Control SM)*

So some teachers felt that it would not give a reflection of what their pupils knew because they weren't used to having to write the answers down on paper. Many teachers mentioned that there are other aspects of learning and demonstrating knowledge that would have been more appropriate were not measured by the assessment.

> *I think the tests don't reflect the children's oracy. When you're in the classroom, I think you can get a lot more from the children whereas the tests, that's limited some of the children in their oracy, that they're able to express through writing, so I think a lot of children are able to access it but a few are not able to express what they truly did understand and what they're doing now. (Full SC)*

Some teachers mentioned that the timing of the assessment and the time frame in which it had to be completed were problematic. They noted that the children's frame of mind was not right for doing that type of exercise just before Christmas and that it was a time of year when a lot of other things were going on.

> *And obviously coming up to Christmas, children's concentration levels and then focusing on Christmas, so in terms of their, their answers to paper and I remember walking round some of them and I know some of them put the wrong answer and I know they could, they knew the right answer, but in their mind frame they were in Christmas mode. (Full SC)*

There were some underlying ethical issues with the pupil assessments in terms of the potential for stress and distress caused to the pupils. For some they were being tested on topics they had not yet covered, in exam conditions they had never experienced and they worried about letting the teacher down, or doing badly in the test. For some pupils this was reported to be quite upsetting.

Difficulty of comparisons

The issues of comparing different cohorts of children and the impact that might have on the results were causes for concern. The idea with an RCT is that overall these differences can be balanced out,

but from the perspective of any one school they could see discrepancies that might influence the data.

> *It's very hard, isn't it? It's very hard to know, it's like with any, anything we test on a cohort of children, you can't run the same thing through on the same cohort of children. (Control SM)*

There was also an issue that the assessments were conducted in very different ways, in different conditions. Some teachers noted that the wording used was not appropriate for pupils in their social context, and that they had to read out questions and 'translate' the wording. Therefore some schools only did a sample of the questions. For example, some said that the wording was very middle class and children in some schools struggled to engage with the idea of "talking about science with mummy and daddy at the dinner table" (Full SC) for instance.

It also became apparent in conversations that many teachers were unclear about which cohorts they were supposed to assess and they were given to different groups of children to those intended in the original RCT model.

> *No, I don't think she has. I think they told her to give them her class, so she's given them her present class. (Partial SM)*

4.3.2 Pupil Attitudes

In terms of pupils' attitudes to science the teachers generally felt they were a good reflection.

> *So I don't know but I mean they would speak honestly and if they thought that they were happy and confident with science they would say, and if they thought that they weren't and they needed to do more I would be quite happy that they would say. So I think it'll be a pretty accurate measurement of their feelings towards science. (Full TC)*

On the other hand, some teachers commented on the possibility of social desirability bias, noting that children might be likely to say what the teachers wanted to hear, especially when they were having questions read to them by the teacher.

> *I just think, not all of the children, some of the children are obviously quite eager to please, and again it's about that type of power relationship that the teacher's asking me to do something and obviously it's been explained to them it's for some people at a university, so it's whether the children are then putting the answers that they think the teachers want. So it is that sort of power relationship so you have to take account. I don't know how you'd measure that really or negate that. It's just what the children do. (Control SC)*

4.4 Perceptions of Involvement

Teachers were asked about how they felt about being involved in the study in general; they talked about their experiences and what they felt they had got out of it, and issues relating to communication and commitment.

4.4.1 Experience of Being in an RCT

For many teachers being involved in the study was a very positive experience, for those in the full intervention group this often referred to the CPD, but teachers in other groups also felt they had had a positive experience.

> *My science coordinator has benefited from being part of this … something like this has helped her develop her management skills. (72 Control SM)*

> *Yeah I think it's quite exciting to – I quite like being part of research; if we get sent surveys I always make sure I fill them in. Yes so I'm quite happy to be part of that. (31 Control TC)*

However, for the control group there were mixed feeling about being involved. Some were happy to be involved in the research, but recognised that you have to do a lot of work for little reward.

> *I think really being part of the research is okay in whatever level. I think the downside of being a participant is you don't know how much work is going to be involved, and if there isn't going to be any CPD immediately that can be a negative thing. It just depends where the school feels they are at the time of the start. (18 Control SM)*

> *I can see absolutely why you'd have a control group, I can see absolutely why you'd do that, it's just hard being in the control group. (38 Control SM)*

Some tried to make the most of the experience, utilizing some of the questionnaire items for their own work.

> *To be honest, being in the control group we didn't really get much from it ourselves. I think it was like when the children did the test and things we... Obviously I looked through them and looked at the pupil voice bit and we did use that across the school for our own evaluation. That was a really nice questionnaire so we picked bits for KS1 and KS2. (87 Control SC)*

4.4.2 Expectations and Commitment

Communicating and explaining the involvement to the teachers was essential to ensure they were committed to the project and that they understood the nature of their involvement.

Communication

Many teachers commented on issues to do with communication and the information they were given. This impacted on recruitment and retention at randomisation and also on the consistency across schools concerning how the RCT was implemented. The control group in particular felt that they were 'out on a limb' and didn't know what was happening.

> *I suppose the hard thing in the sense of there's more contact with the other two schools, as in their knowing more... well obviously they've got tests to be doing and they'll be getting more input there... I suppose you feel like you're out on a limb and everybody else knows what's going on and you don't. I don't know, it's a very difficult one, I think. I think it is just keeping up regular contact and "This is what's going on"*

kind of thing and "This is what'll be happening next" so that everybody knows what's happening. (Control SC)

Some teachers commented that there were happy to do the assessments because they were getting the intervention but that they perceived that many teachers would struggle to keep up this commitment if they were in the control group.

… but I imagine having been on even just the residential, it is a good reminder that we are a really good staff here, because I can't believe how negative a lot of teachers can be with what they have got and the fact that they were there for free, and what they were getting, and still there were grumblings about this, that and the other, so if they had been in the C category, and were having to do the tests and all the rest but not actually getting anything, I imagine probably within a short period of time, your results would have just gone out the window, because they would stop. (Full SC)

The experience of some in the control group led them to say that they would not choose to be in another RCT.

Probably not, to be honest, but I think it's nice to know that afterwards there's going to be some sort of CPD available, but I think everybody who may be involved were all kind of hoping we'd get that full CPD experience, but I've not disliked being part of a project, but obviously as we're in school, you have so many other pressures as well, with my change of roles where everybody's focus shifts a little bit, but no, it's been enjoyable but whether I'd do it again… (18 Control SC)

For some the expectations were more than they felt they had been told at the start. Indeed, it is also potentially problematic that one of the Teacher Colleagues was a teaching assistant rather than a qualified teacher again showing a lack of consistency across schools in the implementation.

Yeah, especially considering I'm a teaching assistant. So, yeah, I think it outweighs it, the expectation is greater than I thought, yeah. That probably wasn't made clear to me to start with anyway. (68 Full TC)

Time commitment

The amount of time they had to spend on the project was also discussed with teachers. For many they felt that the time commitment was fine. This was particularly true for most of the control group. Although for many of the teachers in the full and partial groups it was not an issue either.

It was just another thing I had to do and was part of as well, so it felt like another area of my teaching, but it didn't feel strenuous but it felt that it was just another thing I've got to do as well. (4 Full SC)

On the other hand, for some teachers the experience was perceived to be time consuming. This was almost only among the full intervention group, and tended to refer more to the CPD than the evaluation.

So there were times when I felt I was doing a lot of work and really focusing on Science and you know there are lots of other demands on your time as well. So some – you

know like the residential periods I enjoyed a lot, you know and it was time – it was quite time consuming. I thought it was good; it was good for my subject knowledge but it did take out a lot of – yeah. (34 Full SC)

4.4.3 Incentives

Finally, the teachers were also asked what they might like if they were in the control group as an incentive to stay in the study. Several commented on having money to buy new resources or develop the school in some way. Some teachers requested human resources, such as someone coming in to teach a one-off lesson. Others were very happy with the Amazon vouchers. For some, just the experience of taking part was enough.

For me it goes on my professional record anyway, taking part of it, so it's good for my CV. (72 Control SC)

As one of the recruiters pointed out, at the first stage the freeness of the course was key and even if they were in the control group, getting this eventually was a big incentive.

I didn't tend to ask the Headteachers straightaway, I asked for the bursar or just talked to the secretaries because they tend to know everything that goes on in the schools anyway, and I, you know, emphasised the freeness and the fact that the supply cover was paid for and things like that and I talked about it only being 3% of the country having degrees in Science for primary schools teachers, only 3% of them, and the National was such a prestigious place and that sort of thing basically, and they seemed to like that. That's why I got a lot of recruits. (Julia - Recruiter)

4.5 Validity

In this last section we discuss issues of validity, both from the perspective of the teacher and based on comments they made that might have impacted on the study. The most obvious of these were ways the research itself had impacted on practice, referred to as Hawthorne-type effects.

4.5.1 Hawthorne-type Effects

The impact of being involved in the research influenced teachers in all groups in a number of ways and several themes emerged regarding the extent to which this focus may have affected the data.

Raising the profile of science

Many teachers, particularly in the control group but also some teachers in the partial intervention group, talked about the extent to which being involved in the project had raised the profile of science in the school and how this had affected time and resources dedicated to it.

So it's made sure that we've maintained science in school. We would have anyway, but it's given it an extra edge. (72 Control SM)

I mean, the whole fact that he's part of the study, I mean, we've shared that with the staff and, you know, they're interested in that. They've asked him about, "Oh what are you doing and what kind of things...?" you know, so that's supporting in terms of raising the profile as well. (Partial SM)

For some teachers in the control group, that meant an additional focus on looking for CPD and being driven to do more with science resulting from the disappointment of not being part of the intervention groups.

> *If anything it has made her a lot more driven towards the end result and there's been a massive emphasis on investigative science even down to the CPD and stuff that she's doing in school. (21 Control SM)*

> *I think actually subconsciously it has impacted on us ... and I think you probably created a bit of a "Well, if they're not going to give it to us we'll do something ourselves" … I suppose it wasn't a conscious decision but I think it pushed us to think "Well, we'll show we can do it anyway", so hopefully there has been some improvement. (87 Control SM)*

Many control group teachers discussed the increased emphasis on science and the research prompting them to look more at science, noting that there has been a positive impact despite not receiving the intervention and science was seen as becoming more visible and talked and thought about more in school.

> *It's benefited me because it's actually making me a bit more focused on what we're doing in science and how our results are improving. So even without getting anything from it, it's sort of still going on in the same vein (Control SC)*

Interviews and assessments impacting on understanding

Having people coming into school to talk about science was something that many teachers in the control group said had influenced them and made them more aware of different aspects of science. It prompted teachers to work on areas of need because they felt the pressure of being compared.

> *As I said to you this morning, when you know that somebody's coming in then it does make you evaluate what you're doing. (Control SC)*

For some teachers that resulted in them changing their action plans.

> *Speaking to you is getting me to think about what I need to do and go back and change on my action plan. Just discussing it with somebody, so … (Control SC)*

Doing questionnaires and assessments were also cited as a form of teacher development that had been beneficial in helping teachers recognise where they needed to develop and areas that needed to improve. The assessments were seen as a way to refresh knowledge and think about areas that need revision.

> *It can only be good because I think… because it refreshes things that you don't know about. (Control TC)*

The pupils' assessments were another area that potentially impacted on the practice of the control group. Some teachers were surprised by pupils' responses to questions and used that to highlight areas they needed to address in class. Many teachers found this a very useful process that then impacted on their teaching.

4.5.2 Nature of the Sample and Interfering Factors

The other key issue was that the nature of the sample meant that many of the schools who signed up were already keen on taking science forward and were looking for CPD in order to do this. Many discussed being disappointed because they were at a point when they were ready to focus on science and had this as part of their plan.

> *So we did take some of the resources but obviously we didn't get CPD so we sought our own CPD because that was part of our plan anyway. (87 Control SC)*

The nature of the sample and the different personalities involved was discussed by some teachers who noted that different science leaders would give the research more focus than others.

> *Yes, I think it sounds quite valid, good reasons, and I think having a control as well is good, although I would imagine you'd get variation within the control, wouldn't you, depending on Science Subject Leader and their motivation and inspiration for other people maybe. So I would imagine there'd be quite a lot of variation there. (68 Full SC)*

Teachers from all groups were asked whether they would have sourced more CPD if they had been in the control group and many said that they would not have waited for CPD and would have looked for other input in the meantime.

> *I would have probably looked elsewhere as well, to see what else could supplement it, as well. Because, then, if I was, yeah, I'd definitely, I'd look for some more subject knowledge, yeah, I felt like I needed. (Full SC)*

> *We would no doubt have sought CPD where we felt it was appropriate. (4 Full SM)*

Indeed, some teachers in the full and partial intervention groups had done additional CPD.

> *[Our science coordinator] is one of those people that will go and find other information to supplement it really so in terms of a project we've not really just sat with the, just the Welcome Trust Project, we've taken on other little bits. (85 Partial SM)*

Many of the teachers in the control group had also done additional CPD.

> *We have had our Science leader go off to York to the Science Learning Centre to participate in training with that because we were really disappointed that we didn't get chosen to be one of the Category A schools … So we didn't want to stand still because we thought we could get things moving so we have. (38 Control SM)*

5 Key Themes

In this section we discuss some of the key themes that have emerged from these findings and which inform our understanding of conducting an RCT. These include approaches to recruitment and issues relating to the sample, the scope of what an RCT can detect, communication and development of outcome measures, and the ways the research impacts on participants.

5.1 Challenges of Recruiting and the Nature of the Sample

One of the key lessons learned from this RCT was the difficulty of recruiting teachers for this type research, particularly with the considerable difference between what is received by each of the three groups and the effect of group allocation on the teachers. This ties in with discussions on the nature and size of the sample.

It was difficult to recruit teachers on the research for a number of reasons. First, there were several criteria that the school and teacher had to fulfil in order to qualify for the evaluation. Many teachers were keen to sign up but prevented by having an A-Level or not teaching the right year group; this limited the pool of teachers available. Second, the intervention was very intense and it required people to be prepared to sign up for 24 days of CPD, which many schools did not feel was a priority for them at that time. This had the added implication that those schools that did sign up were those that *did* prioritise science enough to be prepared to do this, which meant that they represented a skewed sample of the whole population. Third, the concept of being randomly allocated to a group put many teachers off. This was an unfamiliar concept and there was quite a lot at stake, which some schools preferred not to risk. Fourth, the evaluation placed a number of demands on the teachers, including completing a subject knowledge assessment, which was daunting for some teachers. They were also required to administer subject knowledge assessments to their pupils and a teacher colleague. Indeed, it is unusual for teachers to be at the same time the subjects and the researchers administering research instruments on behalf of the evaluation team. Moreover, they were also required to ask a colleague to complete a subject knowledge assessment. This gave the evaluation many dimensions and layers.

Given the issues with recruitment and the complexity of the evaluation, splitting the participants into three groups meant that each group was only just large enough to meet the statistical criteria for an RCT. While there was good reason to be interested in the dose effect, it may have been an example of the RCT trying to do too many things as we discuss below.

Allowing enough time for recruitment is essential, and finding ways to help teachers become more familiar with research, and about the use of RCTs in particular, could make a big difference to the recruitment procedure. Being clear about the randomisation, the time and commitment and the benefits and incentives all aid recruitment, and having this done by someone with good knowledge of the target group also helps considerably.

5.2 Combatting attrition

It is important to militate against attrition, and one way is to ensure good communication with the participants, especially at the start; some felt that the time and commitment was undersold. It is also important to think about incentives. Many were happy with Amazon vouchers; others wanted more resources for the school or more guarantees of what they would get in the end. Some teachers, and particularly senior leaders, commented on the idea that schools would appreciate feedback from the assessments and the observations. This would be time consuming to provide, but could be factored into a project where schools valued that input and it might encourage school leaders to get involved.

5.3 The Scope of What an RCT Can Detect

A clear message from this study concerns the nature of the instruments used to gather data in an RCT, and their relationship to the intended outcomes of the intervention. In this case, a subject knowledge assessment was used to measure teachers' (and pupils') understanding of scientific knowledge. In theory, an intervention that aimed to improve teachers' subject knowledge in science could be tested by a random sample of questions from a data bank of validated items (Key Stage 3 Standard Assessment Test items). In practice, no improvements were found in teachers' subject knowledge. This raises the question of appropriate instrument design. It should be well-defined and focus explicitly on the area that requires assessing, rather than trying to detect changes in very broad outcomes.

For this particular study, a decision was made by the funders, evaluators and providers of the CPD that there should be no communication on the topics to be covered in the course and in the assessment between the providers and the evaluators. This resulted in some difficulties with the main research instrument for the RCT. Teachers saw the assessment of subject knowledge very much as a test or examination. However, the nature of examinations is usually that a certain set of knowledge is studied, revised before the examination, and then a subset of that knowledge is tested. This was not the case here and led to many teachers feeling that it did not reflect what they had learned. The 24 days of CPD covered a subset of Key Stage 2 and 3 across physics, chemistry and biology, but the assessments could cover any of that material. Moreover, the teachers did not revise for the assessment and sat it in a range of conditions that were rarely 'exam conditions'. There were also comments that the way they use their knowledge in teaching was that they would look things up on a particular topic before the lesson, making sure they had revised key knowledge; the assessment could not reflect this either. The areas that came up in the subject knowledge assessment in many cases had not been covered in the course, so did not necessarily reflect learning. Yet by having developed more scientific ways of thinking, and learning how to find appropriate information before a class, teachers may still have improved the their ability to teach, even if it did not improve their performance in the assessment.

Similarly for the pupils' assessments, it was often the case that the assessment tested areas that the pupils had not yet covered, so it could not reflect how well that topic had been taught. Depending on the order the curriculum was covered some schools had taught those topics and others had not. This also had implications for the findings.

The broad outcomes articulated for the intervention had an impact on the RCT. For an RCT to be successful it has to address specific and measurable outcomes. The improvement of subject knowledge in general is too broad an outcome to be measured by an RCT. Communication about the areas covered would have given the evaluators the opportunities to focus the subject knowledge assessment. While some evaluators and providers worried this would lead to 'teaching to the test' it seemed that the opposite was the case, with the providers focusing on ways of scientific thinking and overcoming misconceptions in science, which were not picked up by the SATs assessment.

5.4 Communication and Collaborative Definition of Outcomes

Defining what the outcome measures should be, then, is far from straightforward. One lesson that has emerged from this case study is the importance of having a qualitative dimension running

alongside the RCT and there was a general level of consensus that a good evaluation would not rely wholly on quantitative data.

It is also important that there is dialogue between the qualitative and quantitative aspects of the project. There might be merit in gathering some preliminary data to inform the design of the RCT. If the qualitative study had begun before the RCT it would have been possible to use that to inform the development of the RCT, determine what was happening on the ground and devise outcome measures that iron out minor problems before the RCT is set up. While there were timing reasons for wanting to start the project immediately, when investing in research as costly as an RCT it is worth spending the time to ensure these details are worked out, because an RCT can easily tell you nothing.

It would also have been useful to spend that time refining the CPD and determining the key learning outcomes; whether the focus would be on PCK or subject knowledge, the extent to which it would work on thinking scientifically and challenging misconceptions. In generating qualitative data the evaluators would spend more time at the outset with the participants, allowing for deeper understanding of the course and facilitating the development of outcome measures more finely attuned to the course. This would also have meant that the CPD was more fully established before the RCT was set up.

> *One of the problems is that the actual evaluation is taking place at exactly the same time as the design and the delivery of a brand new CPD and when you're trying to recruit schools to undertake something you're trying to recruit them to do an evaluation of something that's totally unknown, that you're still developing. I think that makes it quite difficult because it's a work in process that you're also evaluating. So perhaps it would have been better, in hindsight, to have spent a year with maybe 24 schools or 20 schools looking at the CPD, the type of CPD that might be on offer to iron out any glitches in the CPD delivery in the first place so that then when you can do the evaluation you can say, 'Okay, this is the CPD we've come up on based on teachers' reflections and our reflections in the year. (Sarah - Funder)*

This would also have given time to make informed decisions about who to follow and where the measures could best be taken. For example, the study followed the children across the two years, rather than following the teacher and assessing the teacher's new class in the second year. This was done in some cases, but not with enough schools to generate data for the RCT.

5.5 People as Participants

The influence of the research on the control group and Hawthorne-type effects are inevitable when doing experiments with human participants. The control group can never be blind unless the design uses secondary data. The sample is also likely to be skewed towards those who would be willing to be in the full intervention group, and where that is very intensive, as in this case, that means they are schools that already prioritise science, and thus not necessarily representative. There is also the issue that being part of the research will change people's approach to science. This is difficult to overcome, but ensuring the sample size is big enough to iron out minor differences is essential. It is also necessary to bear these effects in mind when interpreting the findings.

5.6 Fidelity of Data

The final issue was the fidelity with which the assessments were conducted. In some cases teachers did not complete both sets of assessments, or they gave the second round of assessment to a different class. Some teachers took preparation time to complete the assessments, while others did it at home in front of the television. This level of variability could perhaps be ironed out with a large enough sample, but with constant pressure on the number of schools in the RCT, missing data and variable ways of engaging with the assessment could have had an impact.

6 Recommendations

Drawing together some of these discussion points we have four key recommendations regarding conducting RCTs in the future, based on the experience of the RCT in this case study.

6.1 Specificity

Keep the design well-defined and narrowly focused. The qualitative research can gather extra data and look at the causes, but the RCT should measure something that is clearly specified and measurable. Where there is an intervention, the providers and the evaluators need to be clear what each is doing and ensure that the learning outcomes and the outcome measures are closely aligned.

6.2 Simplicity

Keep it simple. Requiring teachers to be both participants and researchers and asking them to make demands on colleagues increases the risk of missing data. The outcome measure should be isolated and simple, with further detail being sought by the qualitative data. Very intensive interventions are more likely to attract a skewed sample as only those who would be prepared to dedicate time to that would be willing to sign up, so the control group is more likely to be pursuing science outside the CPD offered by the intervention and therefore not representative of the total population.

6.3 Scoping

Take time to become familiar with the intervention through qualitative research in the first instance to allow time to use preliminary data to scope the design of the RCT. This would help with the generation of appropriate outcome measures. It could also help with determining the criteria and coming up with incentives in the context of the evaluation in question.

6.4 Sample

Minimise the barriers to recruitment. The sample needs to be big enough to cover missing data and minimise Hawthorne-type effects. This means that there needs to be plenty of time dedicated to recruiting schools and teachers and a large enough pool of possible schools from which to draw. Too many criteria can limit the availability of schools, and schools need to be well informed about the research and about what participating in an RCT means.

We are a global charitable foundation dedicated to achieving extraordinary improvements in human and animal health. We support the brightest minds in biomedical research and the medical humanities. Our breadth of support includes public engagement, education and the application of research to improve health. We are independent of both political and commercial interests.

Wellcome Trust
Gibbs Building215 Euston Road
London NW1 2BE, UK
T +44 (0)20 7611 8888
F +44 (0)20 7611 8545
E contact@wellcome.ac.uk
www.wellcome.ac.uk

The Wellcome Trust is a charity registered in England and Wales, no. 210183. Its sole trustee is The Wellcome Trust Limited, a company registered in England and Wales, no. 2711000 (whose registered office is at 215 Euston Road, London NW1 2BE, UK).