National College for Teaching & Leadership

Closing the gap: test and learn

Research report
Winter 2016

Richard Churches – Education Development Trust (formerly CfBT Education Trust)

Acknowledgements

Any report such as this would not have been possible without the involvement and contributions of a wide range of people. Firstly, thanks go to the team at Education Development Trust[1] (formerly CfBT Education Trust) for helping pull the report together. In addition, I would like to acknowledge the contributions of the Department for Education analyst team, the Centre for the Use of Research and Evidence in Education (CUREE) who contributed sections related to the training provision on the programme and consultation phases (Philippa Cordingley, Paul Crisp, Natalia Buckler) and University of Oxford (Anne Childs, Ian Menter, Ian Thompson, Nigel Fancourt and Roger Firth) for their sections - adapted for use in the report from their British Education Research Association (BERA) 2015 conference papers. Finally, I would like to thank all of the above for their contributions during the peer review phase of this report and specifically Steve Higgins (Durham University) for his valuable contribution to the recommendations section and observations.

[1] www.educationdevelopmenttrust.com

Contents

List of figures

List of tables

1 Abstract

Closing the gap: test and learn is the first programme in the world to trial multiple interventions simultaneously using a wholly collaborative approach across a large number of schools. Seven interventions were chosen through an extensive and systematic consultation and review of interventions seen as most likely to close the attainment gap for pupils with achievement below the national average in literacy and numeracy.

Collaborative randomised controlled trials (RCTs) were then conducted to evaluate these interventions and four of the interventions were replicated. Alongside this, teachers were trained in a range of research methods. A total of 50 teacher-led experimental studies (including RCTs designed and conducted by schools) were grant funded.

This report describes the programme and its delivery. It also outlines the findings from the large-scale trials, learning from developing teachers' scientific literacy through the school-led research programme and conclusions regarding the efficacy of system-led research approaches such as the ones embedded within the initiative.

2 Introduction

This introduction explains the purpose of the report and the areas that are covered. It then goes on to expand on the purpose of the programme (in the context of government policy) and the wider context in which the programme sought to position itself. Finally, the main innovations that were built into the programme are outlined.

2.1 Purpose of this report

This report explores learning from the process of delivering large-scale collaborative RCTs and from facilitating teachers to deliver their own experimental studies. It includes the following trial results.

Trial results for year 1 of the programme:

- 1stClass@Number (1stClass)
- Growth mindsets
- Inference training
- Numicon Intervention Programme (NIP)
- Response to Intervention – Breakthroughs in Literacy (RTI)

The year 2 replications, focused on pupils eligible for free school meals (FSM), of:

- 1stClass@Number
- Growth mindsets
- NIP
- RTI

In addition, it includes:

- the first trialling of the Research Lesson Study (RLS) approach that was developed by teaching schools with the support of the Centre for the Use of Research and Evidence in Education (CUREE) in year 1
- results of the two years' research into the effectiveness of Achievement for All (AfA) within a teaching school context

As well as these results, the report contains a summary of discussions in three forthcoming journal articles that are being produced by University of Oxford Department of Education (OUDE) which will explore a range of learning from the programme in general. Conference paper versions of these articles were delivered at the British Educational Research Association (BERA) conference in Belfast (15–17 September 2015).

2.1.1 Areas that are covered in this report

The closing the gap: test and learn programme involved far more than the delivery of a group of RCTs. This report seeks to reflect this and discusses several levels of research finding, drawing together quantitative and qualitative research evidence from the 27 months of delivery. Broadly speaking, the three areas covered are:

1 findings related to the programme as a whole and the delivery of collaborative large-scale RCTs

2 the statistical effects of the 7 interventions that were trialled at scale in the 11 separate research projects, including a summary of the 190 inferential analyses – that can be found in full in the technical annex A, section 2

3 learning from facilitating teachers to design, implement and report findings from their own RCTs (and wider forms of experimental research, such as quasi-experimental designs).

2.2 Purpose of the closing the gap: test and learn programme

The requirement to close gaps in attainment for pupils from disadvantaged backgrounds is a high priority for schools. The Ofsted inspection framework[2] asks inspectors to make judgements about the performance of all groups of pupils. In particular, within this, efforts that schools are making to close gaps are scrutinised and schools need to account for their effective use of the pupil premium (PP) grant. Increased funding via the PP grant demonstrates the commitment by the government to ensure the poorest pupils leave school on an equal footing with their peers; and Ofsted inspects schools on this basis.

In addition, the statutory guidance to the new national curriculum, updated in July 2014[3], makes it clear that schools must continue to be rigorous in ensuring all groups of pupils are sufficiently and appropriately challenged. It states:

> 'Teachers should set high expectations for every pupil. They should plan stretching work for pupils whose attainment is significantly above the expected standard. They have an even greater obligation to plan lessons for pupils who have low levels of prior attainment or come from disadvantaged backgrounds. Teachers should use appropriate assessment to set targets which are deliberately ambitious.'

The teaching schools initiative defines six core areas of responsibility for teaching schools:

[2] Ofsted. *Handbook for inspecting schools in England under section 5 of the Education Act 2005* (2015). Available at: www.gov.uk (accessed 19th January 2015).

[3] Department for Education. National curriculum in England: framework for key stages 1 to 4 (2014): 4.1. Available at www.gov.uk (accessed 19th January 2015).

1 School-led initial teacher training

2 Continuing professional development

3 Supporting other schools

4 Identifying and developing leadership potential

5 Specialist leaders of education

6 Research and development.

Against this context, the closing the gap: test and learn programme had several aims:

- to ensure that successful approaches to supporting the academic success of the most disadvantaged children are identified and spread (as appropriate)
- to build stronger links between the teaching profession and universities, helping to develop the academic standing of the teaching profession overall
- to further embed changes so that engagement in research was reinforced as an important part of teachers' practice
- to ensure that teachers were supported and enabled to inform their own practice through the use of robust evidence, with a direct impact on educational outcomes for their pupils
- to complement work supported by the Education Endowment Foundation (EEF) and wider efforts to develop an evidence-informed teaching profession

2.3 The context within which the programme has positioned itself

A forthcoming literature review by Queen's University Belfast shows that over the past 10 years more than 800 RCTs have taken place in an education context from within university departments alone.[4] Alongside this, many other RCTs have been conducted around the world, including the extensive work of the EEF, which has commissioned over 100 trials in the past 4 years.

To date the vast majority of RCTs have been implemented and managed by bodies outside of the school system. This programme sought to test how deeply that engagement could be nested within a school-led approach and learn from that process.

Although a development to be applauded, the sudden expansion of trial evidence presents a potential problem. Paralleling challenges in the health sector and other evidence-based areas of the public sector, the accumulation of evidence is not going to be sufficient in

[4] Professor Paul Connolly, 'The trials of education', BERA conference keynote, Queen's University Belfast, 2015.

itself to transform the use of such research without the direct involvement of practitioners. Hearts as well as minds need to be engaged if behaviours are going to change and evidence begins to inform daily practice. Alongside this sits the question of how to develop scientific literacy within a profession that has to date been largely removed from an understanding of the nature of controlled quantitative research and is therefore lacking in the knowledge, skills and understanding necessary to critique such evidence in order to apply it effectively to daily practice. By contrast in established evidence-based fields, such as medicine and healthcare, the understanding of quantitative controlled research is not only expected but forms a fundamental part of practitioner training and development.

2.4 The main innovations that were embedded within the programme

Against the backdrop above, the closing the gap: test and learn programme sought from the outset to go beyond just the creation of evidence to encourage and explore the potential of collaborative practice in the delivery of large-scale RCTs. To achieve this, a number of innovative strands of activity and engagement were initiated around the 11 large-scale RCTs that were completed. These include the following innovations:

- the programme was delivered within an existing system reform – an initiative which has developed a national network of teaching schools based on the model of teaching hospitals
- an extensive systematic consultation was undertaken prior to the start of the programme in order to identify seven interventions that were believed to close the attainment gap for lower-performing pupils. This process integrated the views of schools themselves with academic assessment and the scrutiny of previous research evidence
- teaching schools themselves recruited trial site schools from within their alliances and associated school networks, encouraging engagement with practitioners and the involvement of a wide group of teachers
- rather than externally intervening to manage and co-ordinate the research protocols around the various interventions, teaching schools appointed a trial co-ordinator who was trained to deliver and manage the trial process and testing procedures across the schools that had been recruited
- teaching schools were given a grant so that they could purchase the interventions their schools were testing from the commercial providers, paralleling the later phases in some pharmaceutical trials. This process allowed for high levels of mundane realism[5]

[5] The inclusion of everyday activities within a research process, to increase its efficacy

- to develop trial co-ordinators' understanding further, the training programme included four half-day training sessions in experimental research methods based on the form of content delivered as part of undergraduate psychology training programmes
- participating schools' understanding of education research methods training was further enhanced by the addition of training days delivered by OUDE covering a range of practical topics from planning research to writing it up in a conference-poster style
- in the second year of the programme it was decided to adapt the delivery in order to include replications of four of the seven interventions.

Finally, following interest and the enthusiasm of some teachers who had attended the experimental research training delivered by CfBT Education Trust (Education Development Trust from 1st January 2016), the National College for Teaching and Leadership (NCTL) made grant funding available for up to 50 schools to undertake their own micro-enquiry projects.

3 The programme structure

Following a more detailed discussion of the innovative and collaborative nature of the closing the gap: test and learn programme, the three phases of programme delivery are explained (consultation, capability and dissemination).

3.1 The innovative and collaborative nature of the programme

From the outset, the programme had an innovative and collaborative structure. The NCTL, an executive agency of the Department for Education (DfE), commissioned an extended partnership to manage the implementation of the programme through a competitive tender. CfBT Education Trust (CfBT) acted as the lead provider for the implementation and dissemination phase of the programme. CfBT worked in partnership with CUREE, OUDE and Durham University. This partnership worked on materials development, training and support for the schools that were involved, including launch events, three training rounds across the country, research development and networking events, online events and contributions to the final report. CUREE and Durham University also led the initial extensive consultation during which over 70 interventions nominated by teaching schools were scrutinised to determine the 7 interventions that would form the heart of the large-scale RCT programme (table 3.1.1).

All teaching schools in England were invited to participate in the scheme and a further recruitment round was conducted prior to the start of the second year and the replication programme. In this second round, teaching schools that had not previously had the opportunity to join were approached and new starters brought on board to increase sample size during the replications.

Table 3.1.1: The roles within the programme partnership

National College for Teaching and Leadership	Leading the programme.
CfBT Education Trust **Centre for the Use of Research and Evidence in Education (CUREE)** **Oxford University** **Durham University**	Materials development, training and support. Training rounds 1, 2 and 3. Networking events. Online events. Contributions to the final reporting**.** Extensive consultation to identify the interventions, led by CUREE and Durham University. Over 70 interventions scrutinised in depth.
Participating teaching schools	All teaching schools were invited to join the scheme. 206 teaching schools

	took part, leading and managing the trials.
Trial co-ordinators	Participating teaching schools appointed a trial co-ordinator to manage the activities of their alliance schools.
Trial site schools	The schools where the large-scale trials were completed.
Intervention training providers	Provided training places on courses covering the interventions for teachers in trial site schools.

In total, over the two-year implementation phase, the programme worked with 206 teaching schools. A total of 673 groups of children in trial site schools completed pre- and post-tests. The role of the teaching schools themselves was also highly innovative, as (following training) they were given the role of directly managing the trials and the testing processes from their schools. Their role included:

- appointing a trial co-ordinator
- recruiting trial site schools from within their alliances and networks
- purchasing places on the training courses provided by the intervention providers
- passing on training and administration details
- managing the fidelity of the trials in the trial site schools, a role with increasing challenge in the year 2 replications, as the schools had to manage separation between a control group and intervention group within the same school.

In the first year, 387 trial sites schools took part in the trials, with 15,292 pupils tested. In the second year, 286 schools (5,530 pupils) completed the trials. Intervention providers maintained their natural roles within the trials ensuring high levels of mundane realism (everyday activity). Although in some cases special events had to be put on to accommodate the volume of closing the gap: test and learn participants, in many cases programme participants attended training alongside other teachers from schools outside the programme who had also purchased the commercial training. Importantly, no additional efforts were made to change the nature of the commercial products beyond the way in which they were normally trained and had been trained prior to inclusion within the

trials. The exceptions to the above were RTI[6], a programme developed in partnership with AfA for an EEF-sponsored trial; and the version of RLS which was written specifically for use on the programme by CUREE. There was, at the time that the programme commenced, no comparable commercial version of RLS available at scale in England. The materials produced by this process were developed as Crown Copyright.

3.1.1 A unique relationship with intervention providers

Unlike many of the other RCTs that have taken place in education, a different approach to the provision of intervention was taken. Where previous trials have tended to own the relationship between the intervention and the schools that are trained in that intervention and then deliver it in the classroom, this responsibility was handed over to schools themselves. Specifically, although schools were given a grant to cover the cost of the training and this grant was administered centrally, it was the schools themselves who were required to contact the commercial provider, book their teachers onto the training (within prescribed windows), and attend the training without DfE/NCTL direct oversight. The same approach was taken with regard to the purchasing of the pre- and post-tests from GL Assessment. The schools also managed the administration of the testing, choosing the date that they would do this (also within a window that was prescribed centrally). This required schools to identify and book suitable facilities to allow the children to take the online test, a function which was challenging for some smaller primary schools that lacked onsite computer facilities at scale.

Adopting this approach meant that the research programme as a whole had a high level of mundane realism, improving the programme's external validity and arguably the generalisability of the findings. In other words, the programme was able to create the type of general conditions that might occur in normal daily life where a school had decided to purchase a place for their teachers on a commercial programme with a view to cascading and implementing that training in their schools – whilst evaluating it with an externally purchased standardised test.

With this, of course, came a possible risk to the internal validity of the trials (particularly with regard to whether the protocol delivery was a fitting representation of the provider's product), as considerable levels of trust were being placed on the schools themselves. Indeed, teacher qualitative evidence supplied during the two end-of-year surveys, and from the national event focus group data, suggested that teachers had adapted interventions to suit their context. Again, however, this is something that would be expected in normal everyday circumstances. There is always a tension in experimental research between external and internal validity. Where a laboratory-style trial can generate high levels of internal validity, its external validity (generalisability of the findings) may be

[6] RTI was originally used with year 6 pupils but was adapted for this programme for use with a broader range of ages.

questioned. The opposite is likely to be true for larger-scale, more extended studies that aim to create real world conditions.

Reflecting the collaborative nature of the project, intervention providers were given the opportunity to contribute their own perspectives through an online survey report. These verbatim reports can be found in the technical annex B and were taken into account in the overall conclusions of the programme, and in one case with regard to the conducting of additional analyses. The reports are presented as they were supplied, with only minor house style, proofreading and privacy-related modifications.

3.2 Programme phases

There were three phases to the programme:

- a consultation phase (January to August 2013)
- a capability phase in which the research programmes and training of schools took place (September 2013 to July 2015)
- a dissemination phase involving an event for early adopters of teacher-led experimental research (as they became known) and a national dissemination event involving focus groups (October and November 2015)

3.3 Consultation phase

The purpose of the consultation phase was to identify a set of interventions which the current evidence supported as being effective in closing the attainment gap for lower-performing pupils, with a view to evaluating them using large-scale RCTs. It was also during the consultation phase that the research design for the first phase of trialling and timelines for pre- and post-testing were determined.

3.3.1 Intervention selection process

The first part of the consultation took the form of online surveys and focus groups with partner teaching schools and schools who had expressed an interest in contributing. A total of 233 responses were received to the survey and 19 teachers were involved in focus groups and discussions. NCTL also asked school leaders to nominate interventions and received 24 suggestions. The survey and focus group data offered a list of over 70 potential interventions. The team from CUREE and Durham University used a process designed to select a long list (of 12–18 interventions) most suitable for trialling on the basis of:

- the assessability of the intervention’s planned outcomes and its suitability for use within an RCT

- the manageability of the intervention within the programme timescales, resource levels and the likely demands on participating schools
- the extent to which the intervention matched the criteria identified by schools in the consultation process and, hence, the likelihood of take-up by them

A four-step process was used to consider these issues in turn:

1. Each intervention was given a high, medium or low grading for each key issue.
2. The interventions with a high grading were then ranked on a prioritisation grid for each issue.
3. Each intervention on the prioritisation grid was given a score out of 9, with the intervention placed at number 1 receiving 9 points, the intervention at number 2 receiving 8, and so on.
4. Steps 1 to 3 were then repeated twice, resulting in a composite score across all three key issues.

The teaching schools' research and development (R&D) advisory group then carried out a ranking exercise for the long-list interventions, considering them in terms of likely take-up and manageability for large-scale trialling. This process, combined with the technical scores identified by the Durham University and CUREE team, produced a provisional final list of interventions which was confirmed as the final list after a number of technical and logistical uncertainties were resolved.

The final list of interventions selected was as follows:

- 1stClass@Number (1stClass)
- Achievement for All (AfA)
- Growth mindsets
- Inference training
- Numicon intervention programme (NIP)
- Research lesson study (RLS)
- Response to intervention: breakthroughs in literacy (RTI)

These are discussed in more detail in section 6.

3.3.2 Systems, processes and guides

The final element of the first phase of the project aimed to help establish the conditions for its successful implementation by the school-based R&D team at NCTL and the phase 2

capability partnership (CfBT supported by OUDE, CUREE and Durham University). This involved:

- establishing (sometimes, negotiating) with the intervention providers the detail of their provision and, in particular, how training would be provided for the trial site schools
- producing broad descriptions of each of the interventions for the benefit of the teaching school leaders who would have the first-line responsibility for co-ordinating the interventions in their participating schools
- devising (drawing extensively on Durham University's expertise) the protocols for managing the interventions to provide the most robust environment for conducting RCTs, given the programme's distributed leadership context
- providing advice and guidance to the NCTL school-based R&D team on other features of the programme, particularly the selection, design and logistics of testing, and the management of the randomisation process
- documenting the process and creating guides and other resources for trainers, teaching school co-ordinators and staff in the trial site schools.

These resources were delivered to the NCTL during July and August 2013, with a final set of documents supplied at the end of August.

3.4 Capability phase

CfBT worked in partnership with three other organisations during the capability phase: CUREE, who also led the earlier consultation stage, OUDE and Durham University.

There were three elements to the capability phase of the programme:

1 the delivery of comprehensive training for teaching schools

2 the provision of support to teaching school trial co-ordinators

3 teaching school testing and intervention delivery within trial site schools

The delivery of comprehensive training for teaching schools participating in the closing the gap: test and learn programme covered rigorous and robust research methods appropriate for use in schools, including quantitative research methods such as RCTs, so that teachers gained an awareness of research methodologies (set-up, design and evaluation) and were able to contribute effectively to the trials. This also ensured that teachers in different contexts were able to deliver the interventions under trial in a consistent manner. The strand of work delivered through the RDNE events focused on training teachers in the delivery of small-scale RCTs (and other forms of experimental research) and immediately yielded school-level activity. In response to this, the NCTL made available 50 'early adopter' grants to support participating teaching schools and their alliances in delivering

their own small-scale RCTs. A total of 48 of these studies were presented at a conference poster event at NCTL in Nottingham on 21 October 2015.

All materials supplied to teaching schools were presented in a format that supports cascading and re-delivery to teachers in the wider body of trial site schools. For example, the launch event and training round 1 materials were included on a CD-ROM and teaching schools were supplied with a binder to help them organise the materials. As new materials were delivered these were all made available in an online 'Dropbox' which included a video of the first launch event to support schools to share the vision and aims of the programme within their networks.

Using the materials supplied at the launch events and during training round 1, participating teaching schools and trial site schools carried out a programme of testing over a period of two academic years (September 2013 – July 2015). This focused on assessing whether the seven selected interventions made a positive difference and whether such effects may be replicable and transferable.

3.5 Dissemination phase

Two dissemination events took place – on October 21 2015 in Nottingham, and on 18 November in London. The first event was attended by teachers who had conducted their own teacher-led RCTs (and other forms of experimental research) ('early adopters') together with some invited guests. The second event was open to all participating teaching schools and included presentations summarising the large-scale trial findings and the small-scale teacher-led studies alongside two teacher research presentations

Both of these events contained focus group sessions, the findings from which are discussed below. The early adopter event was opened by the Chair of NCTL and ended with a summary presentation by a professor from Durham University. The national dissemination was opened with a talk by NCTL's deputy director for teaching schools and school improvement and closed with a panel discussion involving researchers from CUREE, OUDE and Durham University.

Finally, this report was drafted with contributions from CfBT, the OUDE partner team, CUREE, Durham University and NCTL. Large-scale trial results analyses were undertaken by analysts from the DfE.

4 A summary of the support provided to teaching schools

In order to implement the programme a broad range of support was developed. This included launch events, training round training days, research development and networking events (RDNEs), materials and a helpline.

4.1 Training

In relation to the delivery of support in year 1, participating teaching school trial co-ordinators (who were leading the delivery of the research across their nominated trial site schools) were offered attendance at a launch event, two one-day training events (training rounds 1 and 2) and two RDNEs. A further training event (training round 3) was delivered in the second academic year, as well as two further RDNEs.

The content of training events (rounds 2 and 3) and RDNEs (rounds 1–4) included learning about research methods along with expert input that ensured the project remained engaging and developmental for the schools. NCTL provided an online 'hotseats' programme by experts through its online community; however, this was ended on close-down of the member website. This provision and the learning from delivering it is discussed in more detail below.

4.2 Research development and networking events

The four RDNEs, as well as providing the opportunity for participating teaching schools to network and learn from one another's experience, were designed to provide a comprehensive programme of learning that enabled schools to design and deliver their own small-scale RCTs. Thus in turn, they sought to develop teachers' scientific literacy within the context of the programme. The programme, embedded within the four events, covered the following:

- designing an RCT and exploring different research designs. For example, the advantages and disadvantages of between-subject versus within-subject designs, choosing and designing tests to ensure validity and reliability, and pre- and post-test designs and when to use them. Teachers were also taught about quasi-experimental designs and the testing of more than one intervention at once
- implementation, sampling, sample size and randomisation (using Excel). This included managing a trial to avoid confounding variables that might arise as a result of delivery
- statistical analysis and interpretation of findings. This included how to conduct preliminary assumption testing, calculating effect sizes, selecting the right test and reporting levels of significance. CfBT's Excel StatsWizard, which can conduct the

main tests teachers needed to use, was made available during the programme. Teachers with more complex designs were given support by CfBT.

- writing up quantitative research and understanding the conventions that apply to this style of research. This session also sought to develop schools' capacity to critique such research and included the use of poster design, building on the extremely well-received training delivered by OUDE at training event 2.

A discussion of findings from the delivery of this strand of support can be found in section 8.

4.3 The early adopter of teacher-led randomised controlled trials grants programme

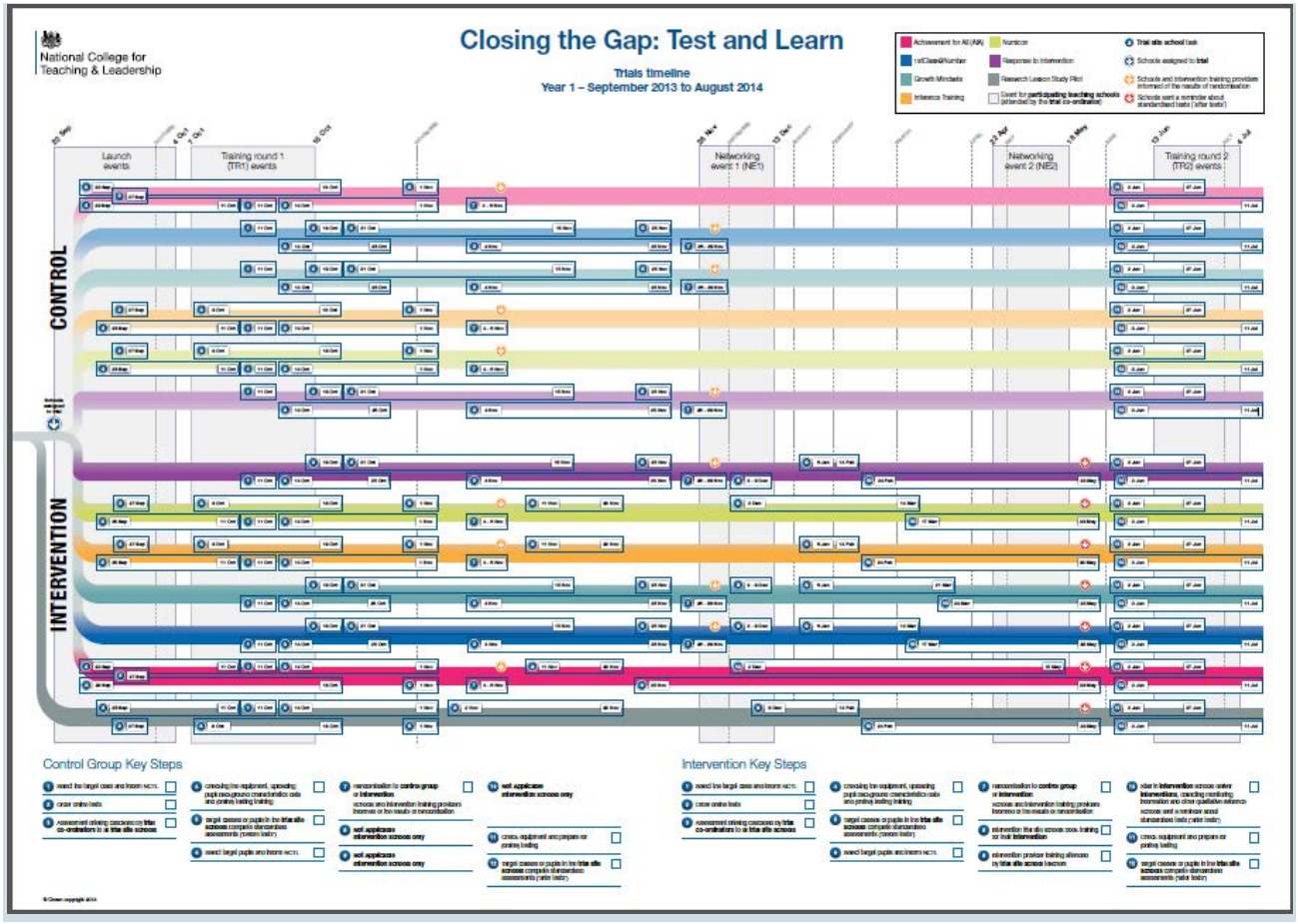
The early adopter grants programme developed from the four half-day training modules that were delivered at the RDNEs over the course of the programme. Although these sessions were initially only intended as a means of developing teachers' understanding of experimental research (including RCTs) the significance of the events rapidly became clear. Almost immediately after the first few events, teaching schools began to use the materials and to try out experimental forms of research design. Interest in the practical use of the materials increased considerably during the first year of the programme and it was decided to put out a call for research proposals.

Fifty grant awards were made available. In total, 66 high quality applications were received. These were evaluated anonymously (and independently) by CfBT and NCTL and the highest-scoring proposals were awarded grants. To further support this strand of delivery, an early adopter launch event took place. This included a series of sessions to help teachers plan further to deliver their designs and build a network of contacts. As this report is being written, 48 of the 50 studies have been completed (see appendix A for the titles of the completed studies).

4.4 Materials

All programme materials designed by the programme were provided in hard-copy format to delegates and made available in an online 'Dropbox' access point. A1 trials timeline posters outlining key dates during the first and second years of the programme were designed, produced and circulated to trial co-ordinators (figure 4.4.1). A closing the gap: test and learn information leaflet aimed at informing higher educational institutes (HEIs) about the programme was developed and disseminated via participating teaching schools. In respect of the trialled interventions, the standard materials used by the various commercial suppliers were given to schools in the interventions as part of the payment made by the schools directly to those suppliers.

Figure 4.4.1: Example trial timeline map



4.5 Helpline

A dedicated email address and telephone helpline for use by participating teaching schools has been operational since the beginning of the programme. A word cloud representing the types of query that were received and dealt with during the programme can be found in figure 4.5.1. In total during this period, the helpline fielded 690 telephone calls and 1,250 email queries.

As the cloud shows, the vast majority of queries were the result of participants needing support with test ordering, the implementation of the standardised tests or the test timelines. The next most common set of queries came in response to the need for help with the funding processes within the programme, the intervention provider training programmes and the training delivered from the programme itself

Figure 4.5.1: Word cloud showing support given by the helpline

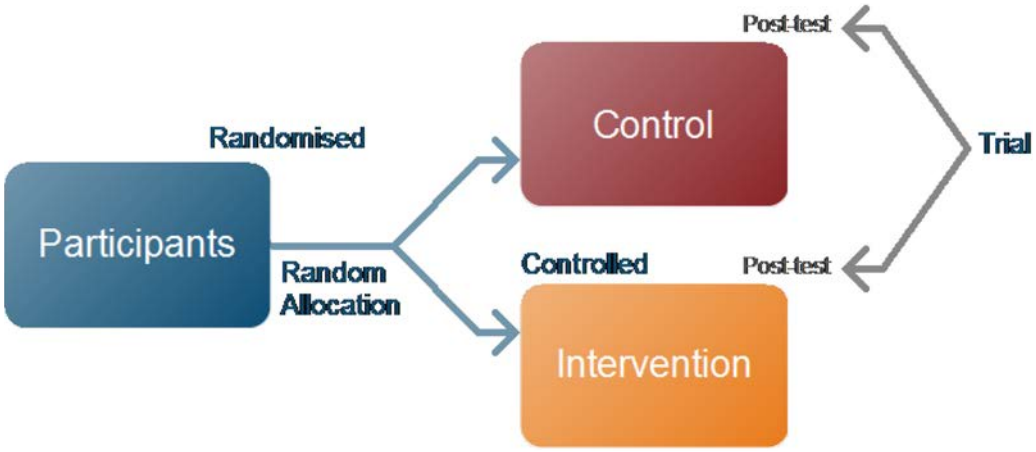
.

5 Randomised controlled trials – concepts and terminology

5.1 What is a randomised controlled trial?

An RCT is different from other forms of research in that a control group is introduced in order to remove biases, such as the fact that an intervention group might have improved even without experiencing the treatment. In addition, for a study to be classed as an RCT, some form of random allocation needs to take place so that the researcher does not directly choose which participants experience either the control group or the intervention group. Finally, there needs to be measurement at the end of the process – the 'trial' (see figure 5.1.1).

Figure 5.1.1: The simplest form of randomised controlled trial



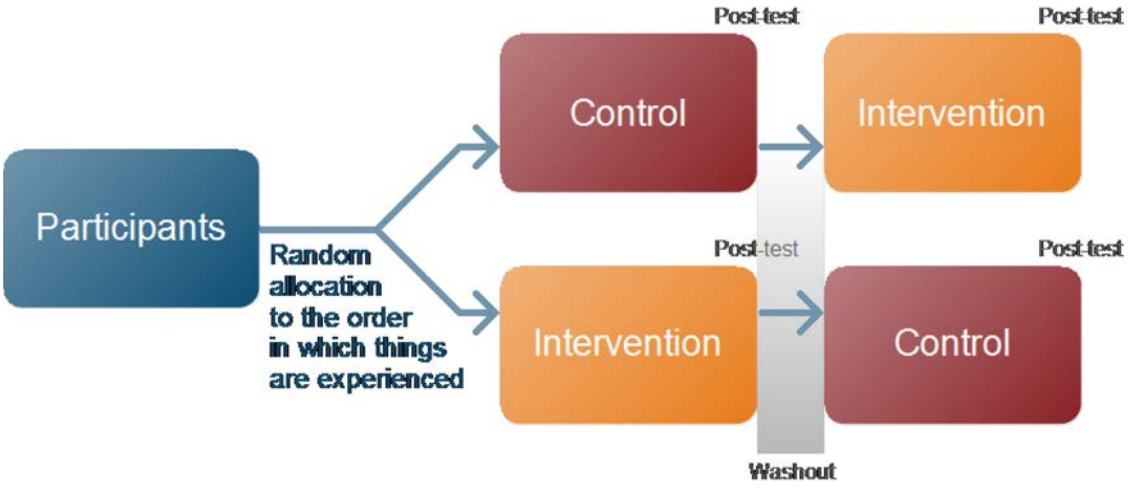
The form of measurement used could be qualitative or quantitative. In most cases RCTs are associated with quantitative measurement and statistical approaches that seek to infer how likely a result may be to generalise to the population from which the sample has been drawn. In other words, the extent to which any change that has been detected may have occurred by chance. This is expressed using a statistic known as the p-value (e.g. p = 0.03, the probability that the result might occur three in a hundred times). For most research, by convention, a level of 0.05 is set as the threshold below which a finding is considered to be 'significant'. In this way, a finding from a study which produced a p-level of p = 0.02 would be considered significant whereas a finding of p = 0.78 would not. This threshold may be made more challenging for studies that are highly invasive or which involve a degree of risk to the participants (as in many clinical trials).

An important point arises from taking such an approach. As such findings are in essence descriptions of probability there is always the chance that results from a single trial may not represent the average effects of a treatment in the real world. This is the reason that replication (the repetition of a piece of research) is essential for a full understanding of a treatment to be developed over time. It is also the reason that it is essential that non-significant results are fully reported as well as significant ones.

5.1.1 Types of design that are suitable for use in a randomised controlled trial

There are various research designs that can be used in an RCT. These include between-subject (or 'independent measures') designs, where half the participants experience the control condition and half the intervention (as in figure 5.1.1). Alternatively, they can be within-subject (or 'repeated measures') designs, where people experience all the conditions, with these presented in different orders to mitigate against any effects carrying over from the control to the intervention and vice versa (figure 5.1.2).

Figure 5.1.2: A within-subject design

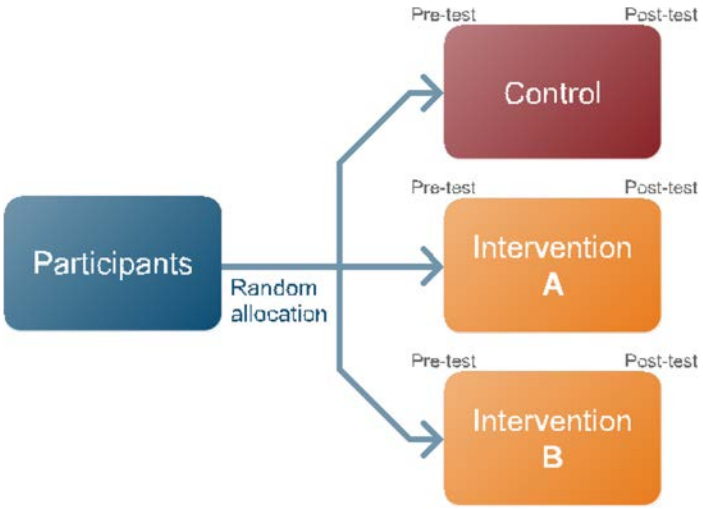


Within-subject designs are usually referred to in medicine as 'cross-over' trials and in psychology often as 'repeated measures' designs. Within-subject designs may include a washout period between conditions, to allow effects to wear off.

In addition, RCTs can make use of case-matching (a process by which participants are paired according to pre-existing characteristics and symptoms before each member of a pair is randomly allocated to either the control or intervention) – a matched pair design. It is also acceptable to test more than two things at once, for example by adding a third condition. This could be a condition that has been designed to control for some aspect of the research to enhance the interpretation of the findings, or it could be another intervention. Adding a second intervention makes the trial more efficient, as two interventions can then be trialled against one control (figure 5.1.3). However, it may not be practical to do this.

Blocked randomisation (or stratified randomisation) can also be applied to reduce participant differences between groups by ensuring a balance of individual characteristics that it may be considered important to control for (such as gender or past experiences).

Figure 5.1.3: A randomised controlled trial with three conditions assessed at once



5.2 Advantages and disadvantages of different designs

There is no correct form of RCT. All of the types of research design above have advantages and disadvantages which challenge the researcher with regard to the validity (being fit for purpose) and reliability (consistency) of the findings. For example, where the between-subject design is useful in situations where the effects of a treatment will not wear off and therefore participants cannot then complete a control condition afterwards (as in the within-subject design), such designs come with challenges when it comes to dealing with between-participant variation. At the end of the day, if you use a between-subject design your results could always be the result of individual differences inherent in the two groups you have compared. Applying more sophisticated forms of randomisation, such as blocked randomisation (which enables the balancing of participant difference between groups) or case-matching (where each participant is paired with a similar other, prior to randomisation) can help. However, individual differences can never be fully eradicated. The advantages and disadvantages of between-subject designs compared to within-subject designs can be summarised as follows.

Table 5.2.1: Advantages and disadvantages of between-subject (or independent measures) designs

Advantages	Disadvantages
Can be used when the effects of a treatment are irreversible	Needs a larger number of participants for an effect to be detected as being significant
Reduces the chance of participants becoming bored by experiencing more than one condition and multiple testing	Variability between participants can affect the results

Advantages	Disadvantages
Removes the risk of becoming better simply through practice	Following randomisation there may be differences between the control and intervention groups which need to be accounted for in the analysis

Table 5.2.2: Advantages and disadvantages of within-subject (or repeated measures) designs

Advantages	Disadvantages
Requires fewer participants (in the case of a design with two conditions, fewer than half)	Participants may experience fatigue particularly if there is a pre- and post-test at the beginning and end of each repeated condition
Reduces the error associated with individual differences as each participant is essentially acting as their own baseline control	Results may be influenced by order or carry-over effects. In other words, results from the second condition may be influenced by the first as every individual participant will have experienced everything within the trial
Small sample sizes can produce valid and reliable results	

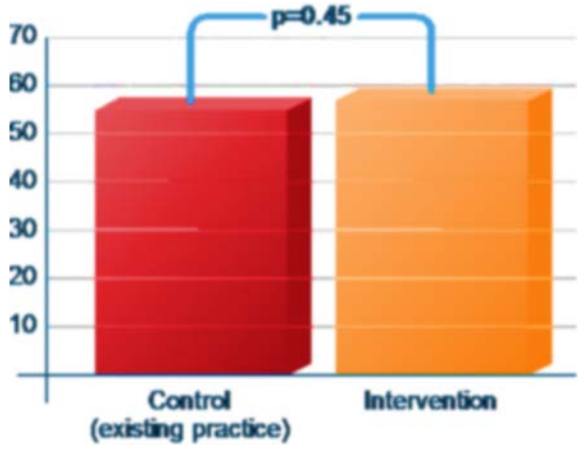
5.3 Randomisation and the content of control conditions

Random allocation can be conducted at a range of different levels. Individual pupils, classes and even whole schools, can be randomly allocated. Where random allocation takes place at a level other than the individual participant it is called a cluster randomisation, with the research design referred to as a cluster RCT. It is usually practical considerations within the context that determine at which level randomisation takes place. Cluster randomisation comes with the inevitable caveat that between-participant variations may have been obscured as a result of having randomised participants in groups.

As in many clinical trials, in the vast majority of cases the appropriate form of control condition is not the removal of education (doing nothing) but rather the use of existing current best practice. As well as the obvious ethical issue that could arise from withdrawing education from a group of children, it would be futile to compare no teaching to some teaching, since all teaching is likely to have some effect on attainment or

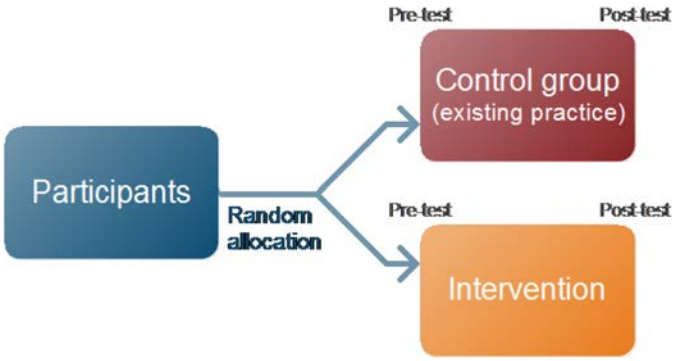
progress. The use of an existing treatment as a control condition is often referred to as a positive control, the use of no treatment at all a negative control. Deploying a positive control consisting of existing current practice changes the way that results need to be interpreted. Specifically, finding a non-significant result (see figure 5.3.1) in this context needs to be interpreted as having identified an alternative treatment that does no more harm that existing practice. Of course, other considerations then need to be applied to such a result, such as the relative cost of the intervention and any other negative side effect of changing to the new treatment (such as teacher workload or the invasiveness of the approach).

Figure 5.3.1: Identification of an alternative treatment



5.4 The pre- and post-test between-subject design

Figure 5.4.1: A between-subject pre- and post-test randomised controlled trial



The most common form of RCT used in education so far is the between-subject pre- and post-test design (illustrated in figure 5.4.1). This type of design is common because many of the interventions that researchers seek to test in education are irreversible and therefore the within-subject design is impossible. The addition of a pre-test, although unable to mitigate against the risk of between-participant variation, can help with interpretation of the final results by identifying situations where there was a big difference between the control and intervention groups – situations that are not uncommon and which then require different forms of analysis.

6 Closing the gap: test and learn – the large-scale trials

Six of the seven interventions trialled were existing programmes, owned, licensed, or managed by charities supported by chargeable training courses[7]. The exception to this was RLS, which was designed specifically for the programme by CUREE, based on the practice of lesson study that was developed in Japan. Whereas all the other interventions were immediately available, enabling participants to purchase places on the training programme within a specified date window (as indicated on the A1 co-ordination map (figure 4.4.1)), RLS needed to be adapted for the purposes of the programme. The developed protocol was then piloted in year 1 of the programme with 20 schools.

6.1 The interventions that were evaluated

6.1.1 Achievement for All

AfA is a whole-school improvement framework[8] lasting two years focused on the lowest achieveing 20 per cent of learners. It was developed and is delivered by Achievement for All 3As. The intervention works through four dimensions: leadership of achievement for all, teaching and learning, parental engagement and wider outcomes. AfA was evaluated using an RCT over two years.

6.1.2 1stClass@Number

1stClass@Number is delivered by trained teaching assistants (TAs) to small groups of pupils in year 3 who have fallen behind in mathematics. The intervention was designed and is delivered by Edge Hill University[9]. TAs work with pupils for eight weeks using detailed lesson plans and adapting them according to information gained from structured assessments.

6.1.3 Growth Mindsets

Growth mindsets is a training programme developed by the University of Portsmouth[10]. It uses approaches to teaching and learning aimed at creating 'growth mindsets' developed

[7] All but RTI and RLS were commercially available. RTI was Crown Copyright (via EEF).

[8] Link to AfA website

[9] Link to Every Child Counts website

[10] Link to University of Portsmouth website

from the research by Carol Dweck[11] which indicates that teachers' and students' beliefs about intelligence have an impact on learning.

6.1.4 Inference Training

Inference training was developed by Leicester City Council[12] based on Yuill and Oakhill's research[13]. It claims to help students make meaning as they read. This involves learning vocabulary, using their background knowledge, making inferences and building up meaning.

6.1.5 Numicon Intervention Programme (NIP)

The NIP approach[14] develops conceptual understanding in mathematics using multi-disciplinary/multi-sensory approaches, making use of apparatus and focusing on action, imagery and conversation. NIP is normally aimed at year 2 pupils working below age-related expectations.

6.1.6 Research Lesson Study

A version of RLS was developed in partnership with schools during the first year of the programme by CUREE[15]. This was trialled in year 2 of the programme. RLS is a structured professional development process in which teachers systematically examine their practice and work together to improve it. Teachers worked collaboratively on a small number of 'study lessons', in a plan-teach-observe-critique cycle. To provide focus and direction to this work, teachers selected an overarching goal and related research question that they wanted to explore. The intervention ran for one term and was suitable for early years right through to year 12. The training for participating teachers was one full day and two half days.

6.1.7 Response to Intervention: breakthroughs in literacy

RTI is a multi-tier approach to the early identification and support of targeted pupils from years 5 to 8 with learning and behaviour needs who are not achieving the age-expected

[11] Blackwell, S., Trzesniewski, K.H. and Dweck, C.S. 'Implicit theories of intelligence predict achievement across an adolescent transition: a longitudinal study and an intervention'. *Child development*, 78 (2007): 1: 246-262.

[12] Link to Leicester City Council website

[13] Yuill, N. and Oakhill, J. 'Effects of inference awareness training on poor reading comprehension'. *Applied Cognitive Psychology*, 2 (1988): 33-45.

[14] Link to Oxford University Press website

[15] Link to CUREE website

level in reading and writing. Literacy interventions are selected by teachers on the basis of close case analysis of pupils' reading and writing needs. The intervention is delivered by CUREE[16].

6.2 Research design

A cluster-randomised, between-subject pre- and post-test design (as in figure 5.4.1) was chosen for use within the large-scale RCTs. In addition to the use of a pre-test, the trial design also sought to build a large sample size where possible (another common approach used to reduce between-participant variation).

In the case of the year 1 trials (where schools, rather than individual pupils, were randomly allocated to the control or intervention) stratified randomisation was used. As discussed above, stratified (or blocked) randomisation involves identifying, in advance, factors (or characteristics) in the participant group that may result in an imbalance between the control group and the intervention group participants. Following the identification of these factors, randomisation is modified to ensure that there is an equal balance of these characteristics in the control and intervention groups. Random allocation at school level removed the risk of control group test performance being affected by the intervention.

Learning from the year 1 interim results, four of the interventions that were assessed in year 1 were replicated in year 2. To reduce some of the between-participant variation that resulted in differing pre-test scores, randomisation to control or intervention took place within each school with predetermined groups of children randomly allocated to either the control group or intervention group. Such an approach ensured that there was no longer any between-school variation, as every school had a control group and an intervention group taking part in each of the trials. However, individual pupils were not randomly allocated, meaning that there was a still a degree of clustering within the process.

6.2.1 Ensuring opportunities for year 1 control group schools to access interventions in year 2 of the programme

The programme design incorporated one additional feature, based on the wait list control group design which is sometimes used in clinical studies. Control schools were given enough grant and the opportunity to purchase one of the available interventions in year 2. This is slightly different to the definition of a wait list control group design as the control group schools and specific children within them did not necessarily go on to be exposed to the same treatment for which they they had previously been the control group. In a wait list control group design, participants in the control group go on to be exposed to the intervention that they were compared to in the first phase of the programme. Such designs

[16] Link to CUREE website

are seen as having the ethical advantage of having a control group whilst at the same time allowing participants to also receive treatment.

The provisional analysis of the year 1 results showed that there were sometimes differences in pre-test scores (probably resulting from cluster randomisation at school level). This aspect of the programme design was modified for efficiency and to facilitate replications in which randomisation could be done within each school (with each school having a control and intervention group). In this way, schools that had been 'waiting' to purchase an intervention for use in their schools were not only able to receive the training but were also able to make a contribution by increasing the number of trial results from seven trials to 11 (seven trials plus four replications) (see table 6.2.1). In total 23 new starter schools who began the programme at the start of the academic year 2014/2015 joined the replication groups. Inference Training was not replicated in year 2 because of adverse effect size findings suggested in the preliminary analysis of year 1 results.

In summary, in the case of all of the first time trials, blocked randomisation took place at school level controlling for a number of features (discussed in detail below). The four replications all involved randomisation at a within-school level – in which the schools were asked to identify a balanced A and B group of pupils and then to maintain separation between these so that the intervention materials were not experienced by the control group. Randomisation in all cases was conducted by NCTL, with schools informed of random allocation results following the completion of pre-testing.

Table 6.2.1: Trials, piloting and replication of the seven interventions over the two-year research programme

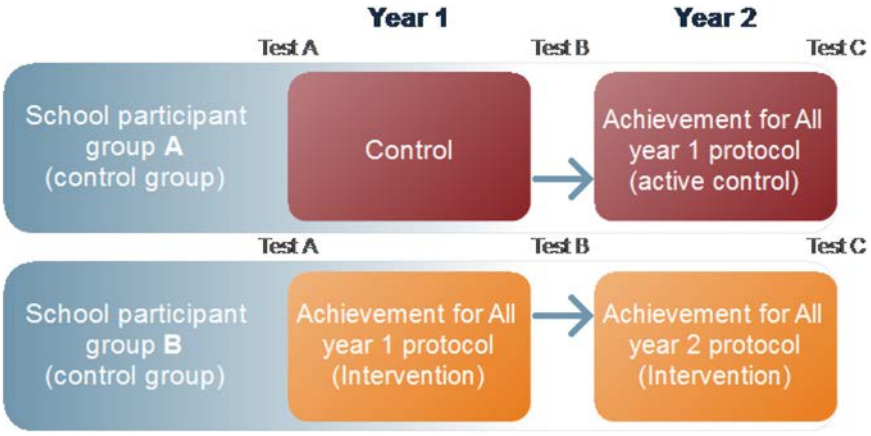
Trial in year 1 with replication in year 2	Trial in year 1 only
1stClass@Number	Inference training
Growth mindsets	
Numicon intervention programme	Piloted in year 1 with trial in year 2
Response to intervention (breakthroughs in literacy)	Research lesson study
Evaluated over two years	
Achievement for all	

6.2.2 Variation in research design structure for Achievement for All

One final variation in the design above was necessary with regard to the evaluation of AfA. AfA is a two-year programme. To ensure that schools who were in the control group for

year 1 had an opportunity to experience AfA in year 2, the following structure was adopted (see figure 6.2.1).

Figure 6.2.1: The adapted design used with Achievement for All



A pre-test (test A) was first taken by all the pupils in both the control and intervention groups, then the control group continued with existing school practice during the first year whilst the intervention group implemented the first year of the AfA protocol. At the end of the first year a 'mid-test' (test B) was taken by all pupils. At this point it was possible to compare the effectiveness of the AfA year 1 protocol with existing teaching school associated practice. This 'mid-test' was then considered to be the pre-test for the second year of AfA. In this second year, the schools that had previously been in the control group moved forward to implement the AfA year 1 protocol, whilst the schools that had previously implemented AfA year 1 now moved to implement the AfA year 2 protocol. The mid-test (test B) then acted as a pre-test for the second year of AfA in which the effectiveness of the first year of AfA could now be compared to the effectiveness of the second year of AfA through the conducting of a final post-test (test C).

When interpreting the results it is important, therefore, for the reader to recognise that where test B allows for an assessment of the first year of AfA against existing teaching school associated practice, test C is assessing the effectiveness of AfA year 1 against AfA year 2, as the control group is now implementing AfA year 1 as an active control condition.

6.3 Participating schools

6.3.1 Recruitment of teaching schools and trial site schools

The scheme was initially advertised through the NCTL member website and the teaching school newsletter. NCTL's regional associates also promoted the scheme to teaching school alliances (TSAs). Teaching schools that expressed interest in the scheme were invited to attend launch events at four locations around England.

6.3.2 The role of teaching schools and trial co-ordinators

Each teaching school nominated schools from within its alliance to take part in individual trials. Teaching school trial co-ordinators were responsible for booking teachers from the intervention trial site schools onto the commercially-available training programmes. Schools were given a grant to cover the costs of paying for this training, cover and travel expenses. Schools also purchased the tests used in the trials directly from the supplier, and were then reimbursed for these costs. No attempt was made to influence the content of the training programmes provided by the suppliers.

6.3.3 Intention to treat and attrition levels

Intention-to-treat analysis is a method of analysing RCTs in which all of the participants who have been randomly allocated to either the control or intervention are analysed together. This is done irrespective of whether they completed or received the treatment or not. Intention to treat is a complex area, and there are a number of competing definitions as to what constitutes intention to treat. Because of the complexity of the present study design with regard to the different stages of recruitment, the point at which intention to treat was seen as commencing was the point at which pupils took the pre-test. We know, however, that in some cases teachers did not necessarily go on to expose all of these pupils to the interventions, a fact which no doubt also contributed to the levels of attrition within the trials (see the technical annex A, section 4).

Pupil attrition rates in the trials where cluster randomisation was conducted at whole-school level ranged from 24 to 48 per cent, with an average of 33 per cent. By far the strongest levels of attrition occurred during the two years of the AfA trial (67 per cent overall). It is likely that this level of attrition was caused by pupil and teacher transition during the long timescales of the trial.

Attrition rates in the year 2 replications, where randomisation was conducted at school level (with a control and intervention group in each school) were much lower. The range of attrition in these trials was between 17 per cent and 66 per cent, with the average attrition rate 23 per cent. One test was severely affected by attrition (the additional writing test for RTI). Removing this test from the attrition results indicates an average of 20 per cent.

6.4 Allocation and randomisation

6.4.1 Phase 1 – preference-based allocation

In the last couple of weeks of the summer term 2013, the participating teaching schools (and their nominated schools) were sent details of the seven interventions and were asked to rank them in their preference order (and to identify any which they specifically did not want to trial). Around half the nominated schools responded to that invitation and these were allocated to either their first or second choice of intervention.

For AfA, the team tried to ensure that only first-choice schools were allocated to it as it was longer (a two-year programme where the others could be completed in around a term), required a higher level of effort on the school's part, and needed to start more promptly than the other interventions.

All remaining unallocated schools (who had expressed preferences) were allocated to the other six interventions – aiming to have a broad spread of schools across all six – using the criteria in the following order:

- first or second choice preferences
- interventions with low numbers of first or second preferences
- phase-specific interventions (NIP and 1stClass@Number)
- all the remaining interventions and schools.

The overall goal was to end up with seven pools of schools allocated to interventions, each of roughly similar size. The exception to this was AfA. Due to the more stringent requirements and higher cost, the final pool for AfA was designed to be slightly smaller than the others.

6.4.2 Phase 2 – random allocation (year 1)

After completing the allocation for all schools that expressed preferences, the programme set about randomly allocating pools of interventions to TSAs so that choices could be made on the ground whilst maintaining the overall size of the intervention pools. Databases were checked for accuracy and duplication and used the resulting master list as the basis for computerised randomisation.

NIP and 1stClass@Number, as phase-specific interventions, were given priority and allocated first. We first filtered the list to include only primary schools and then randomly allocated the required number of schools. Finally, the remaining interventions were allocated, randomly, to bring them up to the same pool size as NIP and 1stClass@Number. Trial site schools were blocked by TSA and then randomly allocated to control or intervention using the RAND() function in Excel. The blocking ensured a balance of control and intervention schools by geographical location and phase.

6.4.3 Phase 3 – random allocation (year 2 replications)

The schools taking part in the RLS trial were randomised in the same way as the schools that took part in the trials in year 1. For the year 2 replicated trials, A and B groups of pupils in each school were randomly allocated to control or intervention using the RAND() function in Excel. Simple randomisation was conducted without stratification.

6.5 Research ethics and the large-scale trials

The closing the gap: test and learn project raised various ethical considerations and challenges at the planning stage, during the trials, and during the analysis. They were broadly reviewed in the light of the BERA guidelines[17], although the involvement of several intervention providers, nearly 900 schools (if teaching schools and trial sites school numbers are considered together) and over 20,000 pupils clearly raised some practical issues.

Initial concerns were around voluntary informed consent, confidentiality and anonymity. The most rigorous position would be for explicit parental and individual consent for every pupil involved. However, it was decided that consent from the headteacher would be appropriate, given that under current legislation teachers have some legal responsibility for pupils' welfare anyway, that schools could choose to adopt the interventions without parental consent, and that the main sponsor was the DfE (through NCTL), which already handled attainment data on all pupils. This decision did not undermine the UN Convention on the rights of the child, to act in the best interests of the child[18]. Further, whilst the project aimed at addressing the needs of educationally disadvantaged pupils, this group were not so vulnerable that special consent was needed, especially as their data was often nested within whole class data.

Pupil data was stored securely by the research team at NCTL on an encrypted and password-protected basis. At the analysis stage, no pupil, school or alliance has been identified. However, as schools and alliances have become involved in further projects, such as 'early adopters', their involvement has voluntarily become public.

The research team was initially concerned that schools would raise ethical concerns about being a control group and not an intervention group, in that it might be argued that pupils were 'missing out'. This was therefore addressed at the initial training sessions, first by pointing out that the interventions should be compared with best existing practice, and indeed encouraging the control group schools to teach as well as they could since the interventions needed to be as good as or better than such practice. Second, a utilitarian argument about the value of system level gains in understanding what works was also proffered. There was no evidence that schools or alliances withdrew from the project on ethical grounds – probably because of the nature of the control condition (existing practice) and the extensive consultation process used to identify the interventions.

The research team was also aware of its ethical obligations to the providers of the interventions. These included treating the relevant data for each provider confidentially, until the research was complete, rather than presenting a partial or distorted picture, for

[17] British Educational Research Association (2011) *Ethical Guidelines for Educational Research*. London: BERA. Link to BERA 2011 report

[18] UNICEF (1989) *The United Nations Convention on the Rights of the Child*. London: UNICEF. Link to UNICEF 1989 report

instance by releasing data or analysis whether positive or negative, before the research was completed. Finally, additional research by OUDE into the project itself (see below) was conducted in accordance with the university's research ethics procedures of consent, anonymity and confidentiality.

6.6 Measures

The large-scale trials used two standardised tests:

- progress in maths (PiM) tests. PiM, developed by GL Assessment and the National Foundation for Educational Research (NFER), covers all current UK national curricula content and assesses mathematical skills and concepts
- new group reading test (NGRT), to measure literacy levels. Developed by GL Assessment and NFER, it assesses the reading and comprehension aspects of literacy.

PiM 6–14 is part of a selection of standardised tests available in both paper and digital editions from GL Assessment. Much of the content of the digital tests is the same as in the paper equivalents. However, some content has been changed to take advantage of digital technology and to eliminate items which did not perform well in digital format. The test questions can be grouped in two ways: curriculum content or mathematical processes. The digital edition of PiM can be administrated to students across the age range of 6–14 years. The purpose of PiM is to enable regular assessment of students, usually carried out once a year (which provides year-on-year progress) and the test content samples the UK curricula.

NGRT is also part of a selection of standardised tests, available in both paper and digital editions from GL Assessment. The NGRT digital edition contains unaltered content from the paper tests but is presented in a way which allows students' reading to be tested according to their performance as they are taking the test, rather than by age or year group, in an adaptive, digital test. The test comprises three sections: phonics, sentence completion and passage comprehension, in two relevant forms (A and B) which can be administered to students across the age range of 7–16 years. The purpose of NGRT is to enable regular assessment of students and can be carried out year-on-year or at the beginning and end of a single academic year as there is an equivalent form. Applying standardised tests like those above (which produce a standard age score), allowed for the combination of different age groups within each trial (see the technical annex A, section 3).

6.7 Hypotheses and analytical approaches

The nature of cluster RCTs inevitably produces results which are harder to interpret than studies with more tightly controlled randomisation. Such challenges are further amplified as the length of treatment period increases and variations in the populations being studied

become increasingly likely to attenuate any effects. This is almost certainly the reason why larger, longer-term education trials tend to produce smaller effect sizes and may be less likely to achieve levels of significance.

Taking this into account, a number of different analyses were conducted on the trial data in order to evaluate six distinct but related hypotheses (table 6.7.1). Separate analyses were conducted on all pupil data and on a sub-group of pupils who were eligible for FSM at the point that they took the pre-test. One of the trials (growth mindsets – year 2 replication) involved the assessment of pupil data from both NGRT and PiM. The RTI year 2 replication was assessed using an additional writing test as well as NGRT.

Table 6.7.1: Hypotheses

Experimental hypothesis 1a	Adjusting for pre-test scores, there will be an improvement in post-test scores for pupils exposed to the intervention for: (i) all pupils and (ii) FSM pupils
Experimental hypothesis 1b	Adjusting for both the design effect caused by cluster randomisation and pre-test scores, there will be an improvement in post-test scores for pupils exposed to the intervention for: (i) all pupils and (ii) FSM pupils
Experimental hypothesis 2a	There will be an improvement in the progress rates of pupils exposed to the intervention for: (i) all pupils and (ii) FSM pupils
Experimental hypothesis 2b	Adjusting for the design effect caused by cluster randomisation, there will be an improvement in the progress rates of pupils exposed to the intervention for: (i) all pupils and (ii) FSM pupils
Experimental hypothesis 3a	There will be a relationship between exposure to intervention and post-intervention test scores, taking into account pre-test scores, gender, age, FSM status, school Ofsted band, and proportion of FSM pupils in the school for: (i) all pupils and (ii) FSM pupils
Experimental hypothesis 3b	There will be a relationship between exposure to intervention and post-intervention test scores, taking into account pre-test scores, gender, age, FSM status, school Ofsted band, proportion of FSM pupils in the school and clustering of participants for: (i) all pupils and (ii) FSM pupils

6.7.1 Preliminary assumption testing and the inferential tests used

Prior to analysis, the principle was adopted that it is not acceptable to drop an outlier, just because it is an outlier – in order to use a parametric test. There was no evidence to suggest that any outliers were caused by data error, deliberate or accidental misreporting, sampling error or the non-maintenance of the research protocol.

To ensure that the correct statistical test is used, a number of assumptions about the data need to be tested. Further details of the assumption testing is available in the technical annex A, section 1.1.

A non-parametric form of ANCOVA (Quade's F)[19] was used to test hypotheses 1a (i–ii) and 1b (i–ii).

To test hypotheses 2a (i–ii) and 2b (i–ii), gain scores were first calculated from pre- and post-test scores and the Kruskal-Wallis one-way ANOVA was applied.

Finally, in relation to hypotheses 3a (i–ii) and 3b (i–ii) the data was evaluated according to the requirements for conducting regression modelling with standard multiple regression. The majority of the data were satisfactory; however, two of the analyses had sample sizes that were too small for effective regression modelling to take place.

Using this variety of tests meant that the first two assessments were able to consider the effectiveness of the interventions with regard to the assessment of attainment, the second two relative progress, and the final pair of assessments were able to provide a validation of the other findings with regard to the effect of a range of individual differences at cluster and pupil level.

6.7.2 Adjustment for design effect caused by cluster randomisation

Adjustment for the design effect resulting from clustering in the testing of hypotheses 1b (i–ii) and 2b (i–ii) was carried out using the approach recommended and discussed by Campbell and colleagues[20]. This approach has been applied to primary healthcare where randomisation has had to take place at GP surgery level. It involves using a formula to calculate the effective sample size taking into account the number of clusters and intracluster dependence. The p-value for the result is then adjusted accordingly. Hypotheses 1a (i–ii) and 2a (i–ii) form an assessment of the effect of interventions without taking into account this design effect whilst hypotheses 1b (i–ii) and 2b (i–ii) represent adjustments to the results found in the preceding hypotheses. Interpretation of these two sets of results allowed for an estimation of the effect of cluster randomisation across the whole programme of RCTs.

6.7.3 Target pupils

In the year 1 trials, teachers were asked to identify 'target' pupils (pupils they believed were most in need of having their attainment gap closed). Unfortunately, this data had reliability issues with regard to pupil identification and therefore analysis was not conducted. There were no issues with regard to target pupil data in year 2 replications, as by definition all the pupils involved in these trials were the target group. Teacher

[19] Quade, D. 'Rank analysis of covariance', *Journal of the American Statistical Association* 62 (1967): 1187–1200

[20] Campbell, M.K., Mollison, J., Steen, N., Grimshaw, J.M. and Eccles, M. 'Analysis of cluster randomized trials in primary care: a practical approach', *Family Practice* 17 (2000): 192–196. Available at: http://fampra.oxfordjournals.org/content/17/2/192.long.

identification was not felt to be secure enough for any other intervention data to be re-analysed.

6.8 Results

This section covers four areas:

1. results for all trials related to the six hypotheses above
2. additional analyses exploring the extent to which the different interventions could be said to have closed attainment gaps compared to expected national pupil progress
3. a discussion of limitations and areas that need to be taken into account when interpreting the findings
4. a summary set of conclusions with regard to the efficacy of the seven interventions that were evaluated

Descriptive statistics and inferential test results from the analyses conducted can be found in the technical annex A, section 2. The threshold at which results would be considered to be significant (i.e. unlikely to have occurred by chance) was set at the level normally considered to be an acceptable minimum according to scientific conventions ($p = 0.05$, a five in one hundred probability that the results may have been arrived at by chance)[21]. In all cases the two-tailed probability has been reported. As is considered good practice by many commentators, effect sizes were calculated so that the strength and direction of the effects can be more easily interpreted. Effect sizes (calculated from test statistics) consistent with the tests used and type of distributions are also to be found in the technical annex A, section 2 (r and partial eta-squared (n_p^2)). However, for ease of overall interpretation, all effect sizes have also been converted to Cohen's d.

In the final summary conclusions, converted ANCOVA effect sizes are presented in a combined tables for 'all' pupil data and 'FSM' pupil data. 95 per cent confidence intervals have also been calculated (figures 6.11.1 and 6.11.2). These provide the reader with an interpretation of the likely range of effect size values that might occur in 95 out of 100 repetitions of the research protocols used in the trials and replications. For readers unfamiliar with this form of representation, the longer the whiskers, the less reliable the value. However, such measures are strongly affected by relative sample size and there were large differences in the sample sizes for the different trials. This resulted from a combination of school choice with regard to what interventions they were happy to participate in and attrition. Therefore, the reader should bear this in mind when viewing the summary figures.

[21] Churches, R. and Dommett, E. (2015) *Teacher-led research: designing and implementing randomised controlled trials and other forms of experimental research*, Carmarthen: Crown House.

6.8.1 Adjusting for pre-test scores, did the interventions improve post-test score attainment for pupils exposed to the intervention (hypothesis 1a (i– ii))?

Adjusting for pre-test scores using Quade's non-parametric ANCOVA produced the following findings. The majority of results, both where 'all' pupil and 'FSM' pupil data were assessed, were non-significant suggesting that existing practice in schools that are associated with teaching schools is, in the main, at least as good as the top seven products identified for use within the programme. There were four exceptions to this. These are described below.

In the year 2 replication of NIP, there was a significant moderate positive effect on attainment for 'all' pupils exposed to the intervention ($p < 0.001$) and for 'FSM' pupils within that group ($p = 0.01$). Conversion of these results to Cohen's d produced effect sizes of $d = 0.38$ and $d = 0.51$, respectively.

Four analyses showed negative significant effects on attainment, suggesting that in the case of these interventions and groups, existing teaching school practice was better than the provided alternative. Specifically, the growth mindsets intervention applied in the year 1 trial had a small negative effect on attainment for 'all' pupils ($d = -0.17$, $p < 0.001$) as did the second year of AfA ($d = -0.26$, $p = 0.02$). Moderate significant negative effects were detected with regard to the 'FSM' pupils exposed to AfA during the year 1 trial ($d = -0.43$, $p < 0.001$) and the use of RTI with 'FSM' pupils in the year 2 replication ($d = -0.46$, $p = 0.03$).

6.8.2 Adjusting for pre-test scores and the design effect resulting from cluster randomisation, were the findings from hypothesis 1a supported (hypothesis 1b (i–ii))?

As described above, a design effect was calculated from intra-cluster dependence and cluster size which enable an effective sample size to be calculated and the p-levels for the results in hypothesis 1a to be adjusted accordingly.

Adjusting for cluster randomisation there was no substantial change in the results, which implies that there was little variation at school level – despite concerns that this might be the case, particularly with regard to the year 1 trials where different schools were in the control groups and intervention groups.

In addition, the NIP replication results were still significant for both 'all' pupils exposed to the intervention ($p < 0.001$) and the 'FSM' sub-group of pupils ($p = 0.01$). This was also the case for the negative effects produced in the growth mindsets year 1 trial with 'all' pupils ($p = 0.03$), the RTI FSM replication results ($p = 0.03$), AfA year 2 'all' ($p = 0.04$) and 'FSM' results ($p < 0.002$). All other results remained non-significant and therefore equal to existing practice.

6.8.3 Was there an improvement in the progress rates of pupils who were exposed to the intervention (hypothesis 2a (i–ii))?

Looking at the pupil data with regard to the amount of gain made by pupils in the intervention groups (pre-test minus post-test scores) compared to the control group produced similar results for NIP, growth mindsets and AfA. However, RTI was now found to be equal to existing teaching school associated practice.

There was a moderate significant gain in the year 2 NIP replication for ‘all’ pupils ($d = 0.39$, $p < 0.001$) exposed to the intervention and for ‘FSM’ pupils ($d = 0.51$, $p = 0.01$). In addition, NIP year 1 ‘all’ pupil results indicated a small significant gain in progress ($d = 0.16$, $p = 0.02$).

With regard to negative effects, growth mindsets year 1 ‘all’ results were again significant ($d = -0.13$, $p = 0.009$) as were AfA year 2 ‘All’ ($d = -0.32$, $p = 0.004$) and year 1 ‘FSM’ ($d = -0.50$, $p < 0.001$). All other trials results were non-significant, indicating parity with existing teaching school practice.

6.8.4 Adjusting for the design effect resulting from cluster randomisation, were the findings from hypothesis 2a in relation to pupil progress supported (hypothesis 2b (i–ii))?

Adjusting the gain score results above to take into account the design effect resulting from cluster randomisation changed three of the hypothesis 2a results. AfA ($p = 0.09$), NIP ($p = 0.09$) and growth mindsets year 1 ‘all’ ($p = 0.11$) results were now found to be non-significant. This said, the moderate positive effects on gain detected for NIP year 2 ‘all’ results ($d = 0.39$) and FSM results ($d = 0.51$) remained significant with p-values of < 0.001 and 0.01, respectively.

All other results were non-significant, paralleling the findings in earlier analyses.

6.8.5 Was there a relationship between exposure to the intervention and post-test intervention test scores, taking into account pre-test scores, gender, age, FSM status, school Ofsted band, and the proportion of FSM pupils in the school (hypothesis 3a)?

Regression modelling using standard multiple regression allowed for an exploration of attainment following exposure to the intervention, taking into account a wide range of potentially confounding contextual factors and individual differences. Taking these factors into account, regression modelling supported the findings with regard to the growth mindsets year 1 ‘all’ negative effect size ($p < 0.001$) and the negative effect on attainment that was detected from the AfA year 1 ‘FSM’ data ($p = 0.04$). It also supported the NIP year 2 ‘all’ positive effect size ($p < 0.001$). One difference was that the 1stClass@Number year 1 trial data now appeared to show a positive effect on attainment ($p = 0.01$).

All other results indicated equity with teaching school practice in the control groups.

6.8.6 Adding clustering as a factor within the regression modelling, was there still a relationship between exposure to the intervention and post-test attainment scores (hypothesis 3b)?

Making a further adjustment within the regression modelling by taking into account clustering resulted in hypotheses remaining supported. These were the negative effects of growth mindsets on 'all' pupil attainment observed in the year 1 trial ($p < 0.001$) and the positive effects on attainment for 'all' pupils observed in the NIP year 2 replication ($p < 0.001$). The RTI year 2 'FSM' ($p = 0.18$), 1stClass@Number year 1 'all' ($p = 0.10$) and AfA year 1 'all' ($p = 0.50$) and 'FSM' results (0.06) were no longer significant.

RTI year 2 (writing test), AfA year 2 ('all' and 'FSM') sample sizes were too small for regression modelling to be conducted effectively.

6.8.6 Additional analyses of one set of intervention results following concerns about the fidelity of the trial at school level and assessment of target pupils

One provider, RTI, expressed concern that during the year 2 replication there may have been some cross-contamination – in other words pupils in the control condition may have ended up exposed to the interventions.

Two additional sets of analyses were conducted on the NGRT results: firstly, an analysis removing the schools named by the provider, and secondly an analysis removing only the schools that confirmed there may have been an issue following an online survey of the schools (see the technical annex A, sections 2.24 - 2.27). In all cases, the previously detected negative effects remained consistent: namely, for RTI year 2 'All' pupils d = -0.37 and -0.18 respectively; and for 'FSM' pupils d = -0.36 and -0.14.

6.9 Additional evaluation of the extent to which the pupils' attainment gap (compared to expected pupil progress) had been closed

GL Assessment standard age scores (SAS) are based on the student's raw score adjusted for age and placed on a scale that makes a comparison with a nationally representative sample of UK students of the same age. The national average score is 100. Using the GL Assessment SAS national average score as a baseline, the following charts show how the attainment gaps between FSM pupils and other pupils changed over the period of the trials.

In the trial design used in year 1, schools selected whole classes of pupils for each trial. It was therefore possible to compare mean standard age scores for FSM pupils with mean

standard age scores for pupils who were not eligible for free school meals, in the same way that the government has reported the attainment gap for disadvantaged pupils at the end of key stage 2 and key stage 4[22]. In the design used in year 2, schools only selected pupils that they felt were disadvantaged. In this case, the gap was calculated as the difference between the mean standard age scores for the FSM pupils and the standard mean of 100.

As described above, the AfA trial ran over the full two years of the scheme. In this trial, the intervention group began implementing AfA in the first year, while the control schools carried on as normal. In the second year, the control schools began implementing the first year of the AfA programme, while the intervention schools moved into the second year of the programme.

The bars in the charts show the change in the attainment gap. A positive score indicates that the attainment gap has been reduced for pupils eligible for FSM. A negative score indicates that the attainment gap has increased.

Assessment of the SAS points reduction in attainment gap for the year 1 trials and RLS (figure 6.9.1) shows that in the case of Inference Training, RTI and 1stClass@Number the control condition (existing teaching school associated practice) appeared to have closed attainment gaps more effectively than the interventions. The exceptions to this were growth mindsets, NIP and RLS, which appeared to achieve more relative gap closure than existing practice.

With regard to the first year of the AfA trial (figure 6.9.2), the first year of AfA implementation closed attainment (relatively) far less than existing practice. The year 2 control (which was the repetition of the first year of AfA) continued to produce lower gains in gap closure, although there was a small positive gain achieved for the year 2 AfA protocol.

A number of limitations apply when interpreting the effects represented in the figures below. Specifically, it should be recalled that the interventions were very different with regard to a number of factors:

- although most interventions were well established, others were new (RLS) or had been adapted for use on the programme (RTI)
- there were different delivery periods for some interventions
- the structure and content of the interventions was very different

Figure 6.9.1: SAS points reduction in attainment gaps for the year 1 trials and RLS (trialled for the first time in year 2 of the programme) – control and intervention

[22] Link to social mobility indicators

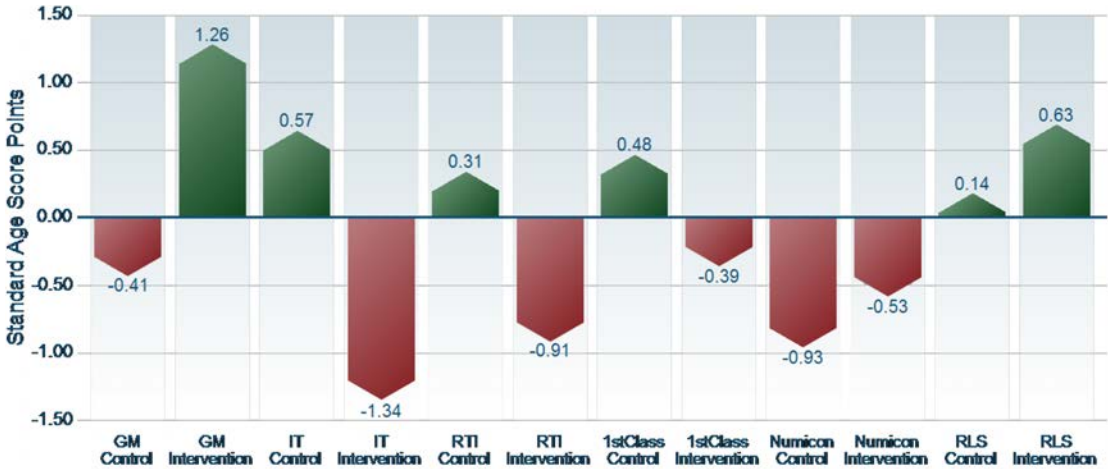
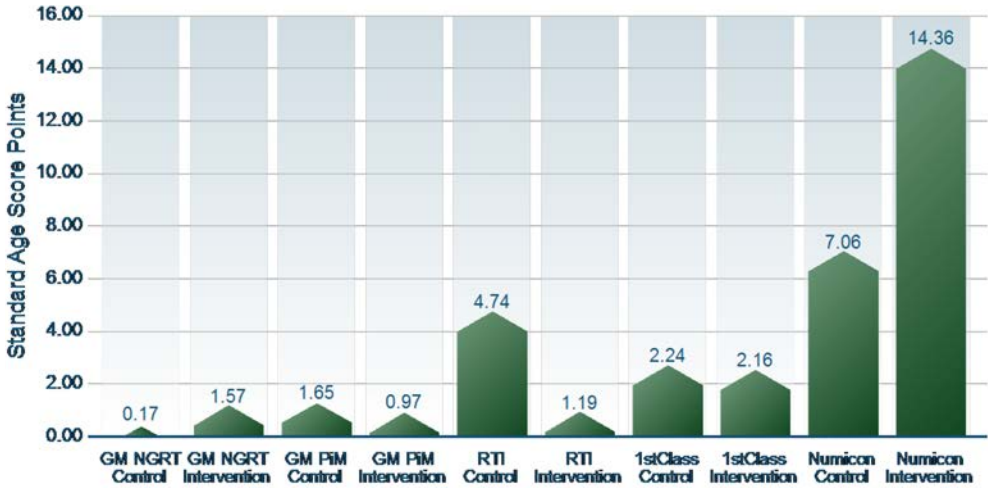


Figure 6.9.2: SAS points reduction in attainment gaps for AfA (year 1 and year 2) – control and intervention

The replicated year 2 trials (figure 6.9.3) yielded positive gains across the board although a similar pattern was evident with RTI and 1stClass@Number which produced less gap closure than existing practice. This was also the case with regard to mathematics attainment in the growth mindsets replication, although growth mindsets appeared to close attainment gaps more effectively with regard to improvement in literacy. NIP, paralleling the inferential test results and effect sizes resulting from the testing of hypotheses 1a–3b, closed the attainment gap of pupils substantially, even compared to a control group that was itself making substantial gains.

Figure 6.9.3: SAS points reduction in attainment gaps for the year 2 replicated trials – control and intervention



6.10 Limitations and important considerations when interpreting the results

As has been discussed earlier, all experimental research has limitations. With regard to the large-scale trials within the closing the gap: test and learn programme, these limitations were mainly the product of cluster randomisation and internal validity issues arising from the collaborative school-led approach which reduced the degree to which extraneous variables could be controlled for. There were also limitations with regard to some of the trials caused by the fact that, to some degree, the use of standardised literacy and numeracy tests (such as when used with growth mindsets and AfA) could be considered to have reduced the design's sensitivity in detecting changes caused by these interventions. Neither growth mindsets nor AfA claim to directly affect mathematics or literacy attainment. At the same time RLS is not a direct literacy approach but rather one which can be tailored for such use. Similarly, RTI is not a direct pedagogical intervention but rather a flexible targeting approach. This meant that although teachers had a choice between targeting the reading or writing skills of their pupils, it was not possible to match the assessment to reflect this level of teacher in-class usage.

It is also important to acknowledge the wide variation in the nature of the interventions which had in some cases very different focuses, breadth and emphases with regard to target pupils. In this respect, the fact that a teacher-identified sub-group of target pupils (in the year 1 trials) lacked reliability and therefore the main analysis reverted to the analysis of effects on the 'FSM' sub-group may have affected some interventions such as RTI. It is also important to remember that some of the interventions had a long delivery history (such as NIP) and have had a far longer time to become established and improve through feedback and revision. In contrast, for example, RLS was a new training programme which was only piloted the previous year, with 20 of the participating schools.

Despite the fact that some trials may have been affected by cluster randomisation, this was by no means the case with the vast majority of trials. Finally, where such between-school effects appear to have occurred, it was not possible to determine the extent to

which these may have been caused by teacher-level differences or school-context differences, as the programme collected no teacher performance data with which to make comparisons.

6.11 Conclusions with regard to the large-scale trials

Pooling the converted Cohen's d effect sizes from hypotheses 1a and 1b produces a map of relative effect sizes as shown in figures 6.11.1 and 6.11.2[23]. Effect sizes range widely (from d = 0.51 to d = -0.43). Interpretation of this combined table, however, needs to bear in mind the wide variety of interventions, different treatment periods and the fact that the confidence interval bars are partly a product of sample size – with there being a wide range of pupil numbers across the trials and sub-groups within the trials.

Within the limitations of the research design and its implementation, the following conclusions can be drawn. Out of around 200 inferential analyses, 25 results were significant (15 positive effects, 10 negative effects). The average effect size across all analyses was d = -0.02. Following concern about cross-contamination in one year 2 replication, results were re-analysed, with schools of concern removed; however, the results remained broadly the same.

Figure 6.11.1: Combined effect sizes (generated from ANCOVA results) for all pupils involved in the trials

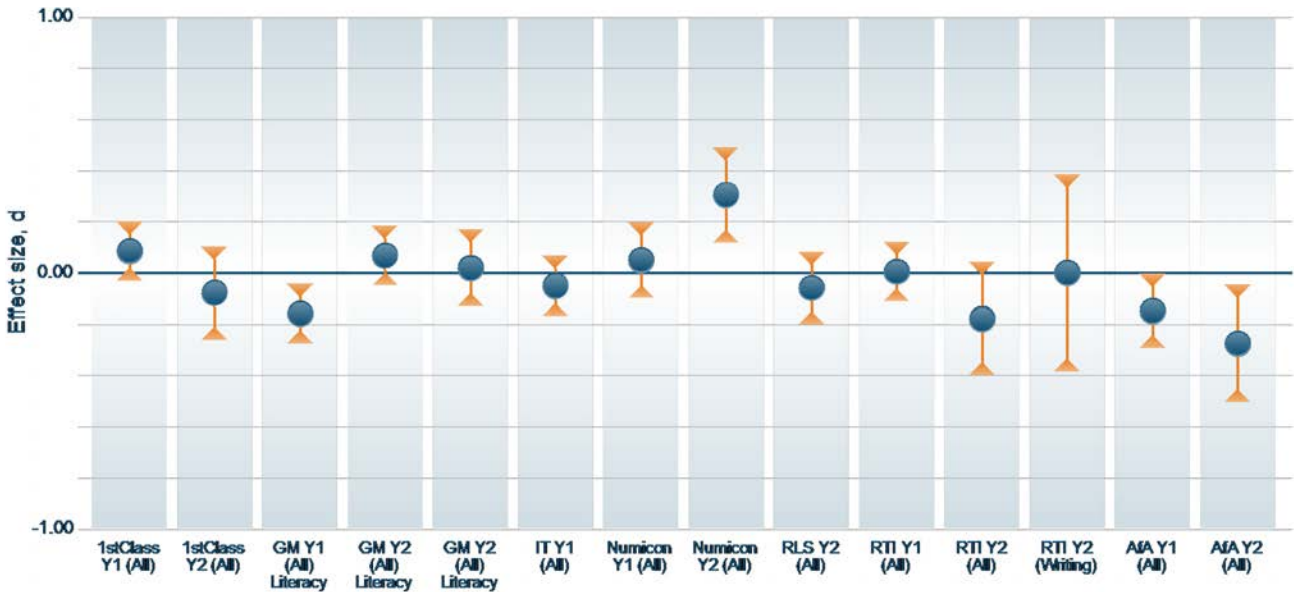
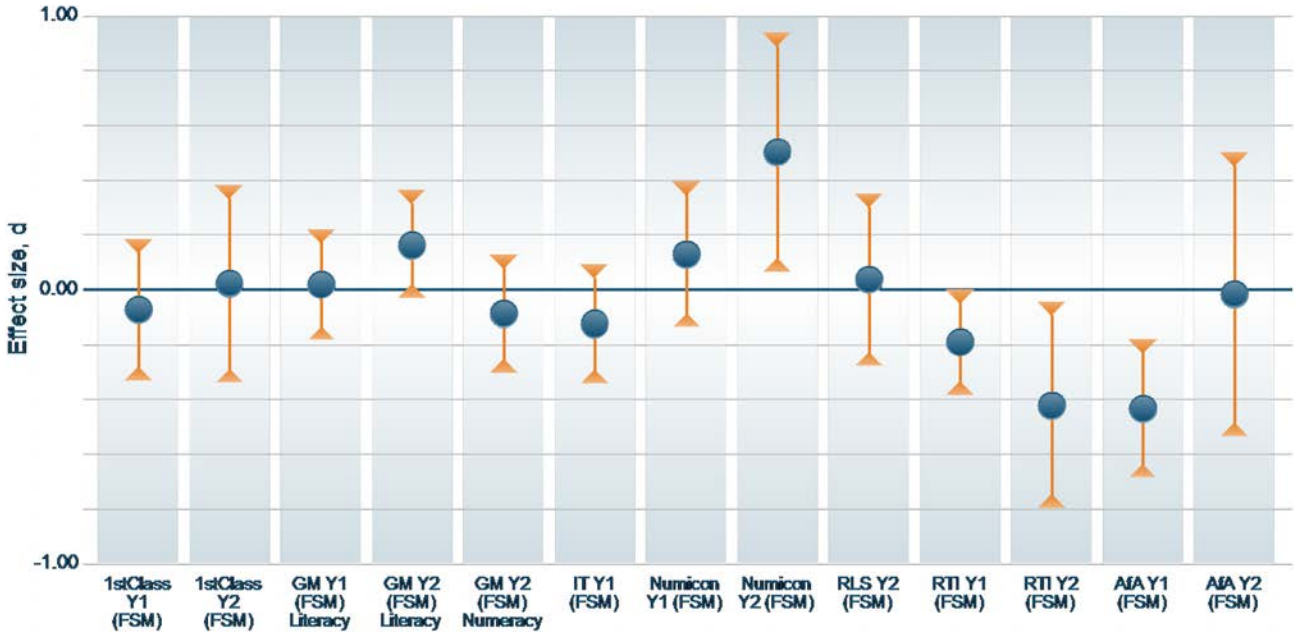


Figure 6.11.2: Combined effect sizes (generated from ANCOVA results) for FSM pupils involved in the trials

[23] The same cautions, noted on page 49 (with respect to figures 6.9.1, 6.9.2 and 6.9.3), should be applied when interpreting these combined graphs.



Some trials may have been affected by cluster randomisation and/or between-school differences, but not all were affected. In particular, AfA results may have been affected by a range of differences. This said, the design did not allow for the assessment of whether school-level differences were at the teacher or pupil level. RTI may have been affected by clustering.

Overall, teaching school associated existing practice (the control conditions for all of the large-scale trials) appears to be at least equal to six of the top seven interventions identified in the consultation but better than growth mindsets when used with an average group of pupils. Existing teaching school practice may also be better than the first year of AfA with regard to the exposure of FSM pupils to this treatment.

The exception to the above was NIP which consistently appeared to improve mathematics attainment and progress rates, particularly for FSM pupils, and irrespective of the analytical model used to assess its efficacy. There may also be gap closure benefits in the use of RLS; however, in the context of the present study design none of the assessments produced significant results.

Although finding that the majority of the interventions showed no effect greater than existing good practice is useful, it is but a starting point for further investigation. The established practice in other fields (e.g. medicine) would be to undertake further trials in different circumstances to see if the benefits of these interventions are revealed in different contexts (for example, in struggling schools, with a more tightly defined group of students, or particular age groups).

7 Innovation and learning within the formal training programme

This section of the report was developed and contributed to by CUREE.

7.1 Accommodating different needs within a coherent structure

The programme was supported by a series of training events, consisting of a launch event and three training rounds for schools leaders and teaching school trial co-ordinators from the participating schools. The training events programme was planned jointly by CUREE and OUDE colleagues to secure coherence between the design and implementation phases. These sessions had to accommodate diverse participants, with a wide variety of levels of experience, both in general and specifically relating to research and development or RCTs. At later events there were also colleagues who were new to the programme. Each component of the training programme aimed to achieve multiple goals, including:

- building commitment to the project
- informing participants about the approach and arrangements
- providing participants (who were trial co-ordinators) with information about the interventions
- (at training rounds 2 and 3 especially) providing more advanced/experienced participants with information about how to carry out their own enquiries

Throughout the training programme priority was given to building on the evidence about effective continuing professional development (CPD) for educators, including the importance of:

- engaging participants in connecting programme approaches with their own experiences and contexts
- focusing on aspirations for pupils
- encouraging structured collaboration
- underpinning the CPD with structured tools to scaffold consistency and make support accessible.

The training events consisted of the following components.

1. A series of launch events, split into the following sessions:

- a brief introduction to the facilitator team
- making target pupils real – an activity designed to relate the notional target pupils for the interventions to real children from their settings, through the use of characters from fiction as metaphors
- a split session consisting of: developing understanding of the goals and key features of the different interventions
- developing an understanding of some of the practicalities of leading participation in the large-scale trials
- understanding the requirements/logistics and benefits of participating in AfA and early engagement with the programme (because of logistical requirements schools allocated to AfA had to be prepared to begin the trial much sooner than others)
- a feedback and reflection session where participants could get answers to specific queries they had supplied earlier in the day (after these queries were collated by facilitators)

2. Training round 1 (TR1) focused on helping trial co-ordinators with the necessary information to fulfil this role via:

- an introductory session, helping participants to connect with the project aims, and explore the different reasons that schools in their networks had for taking part and their different levels of capacity for engaging with trial methodologies
- a recap session to reinforce participants' understanding of the nature and source of the interventions as well as allocations and programme goals
- an exploratory session to develop an understanding of how the trials connected with and reinforced existing R&D
- an explanatory session to give participants an understanding of (and, hopefully, enthusiasm for) RCTs, R&D; and to introduce the role of qualitative evidence in conducting and making use of trials
- a session to develop participants' understanding of the tools and protocols developed to support fidelity in trial implementation, and to enable read-across qualitative work within and between schools and TSAs
- a session explaining the nature of the assessment protocols established for the project
- a practical session exploring the timelines and practical requirements of the interventions which participants would be overseeing across their TSAs

- a question and answer review and summary session

Before the team settled on a structure for training round 2 (TR2), several members conducted phone interviews with R&D leads within TSAs who had been involved in trialling interventions, to gain an insight into what the driving issues for R&D leads were likely to be.

3. TR2 was planned to build on the resulting evidence and consisted of:

- a collaborative session for participants to share progress and experiences so far, and identify goals for continued development
- a planning session in which participants were asked to consider one of two hypothetical scenarios related to helping schools make decisions in the light of RCT results, and make suggestions for how to respond to the challenges these scenarios presented
- an analytical session in which participants developed their own research skills/understanding, through focusing in depth on identifying good research questions and related methods
- a learning session in which participants were introduced to and given a chance to familiarise themselves with a variety of methods of research dissemination
- a reflective plenary session in which participants reviewed the connections between the content of the training and both action planning and evaluation

4 Training round 3 (TR3) took place in the second year of the project, so the design team constructed an offer that was differentiated for ‘old hands’, people in the midst of the process, and those who were completely new to it. TR3 therefore consisted of:

- an introductory session for taking stock of their current situation and identifying development needs
- a session in which participants were given both time and structure for considering what involvement both in TR3 and the closing the gap: test and learn programme as a whole would need to include in order to be a success for them, their schools and their pupils
- a series of parallel sessions

Much was learned from engagement with the trial co-ordinators, many of whom were the teaching schools R&D Lead. Drawing on this and the qualitative data supplied by schools in the annual online survey, the following became clear:

- teaching schools were at very different stages of development with regard to carrying out and participating in a large-scale collaborative research project
- in some cases, the schools were already conducting sophisticated research programmes; in others this was their first research activity – this was usually a function of how recently they had become a teaching school, but not in all cases
- early in the programme some trial co-ordinators had the additional challenge of building a relationship with newly formed or recently formed TSAs – training and helpline support had to be adapted to support this
- providing wider differentiation opportunities was welcomed by teachers with regard to the changes made to the training rounds
- the successful delivery of the programme was underpinned by the sense of belonging that exists within the teaching school cohort and a powerful sense of purpose with regard to being a teaching school and engaging in research.

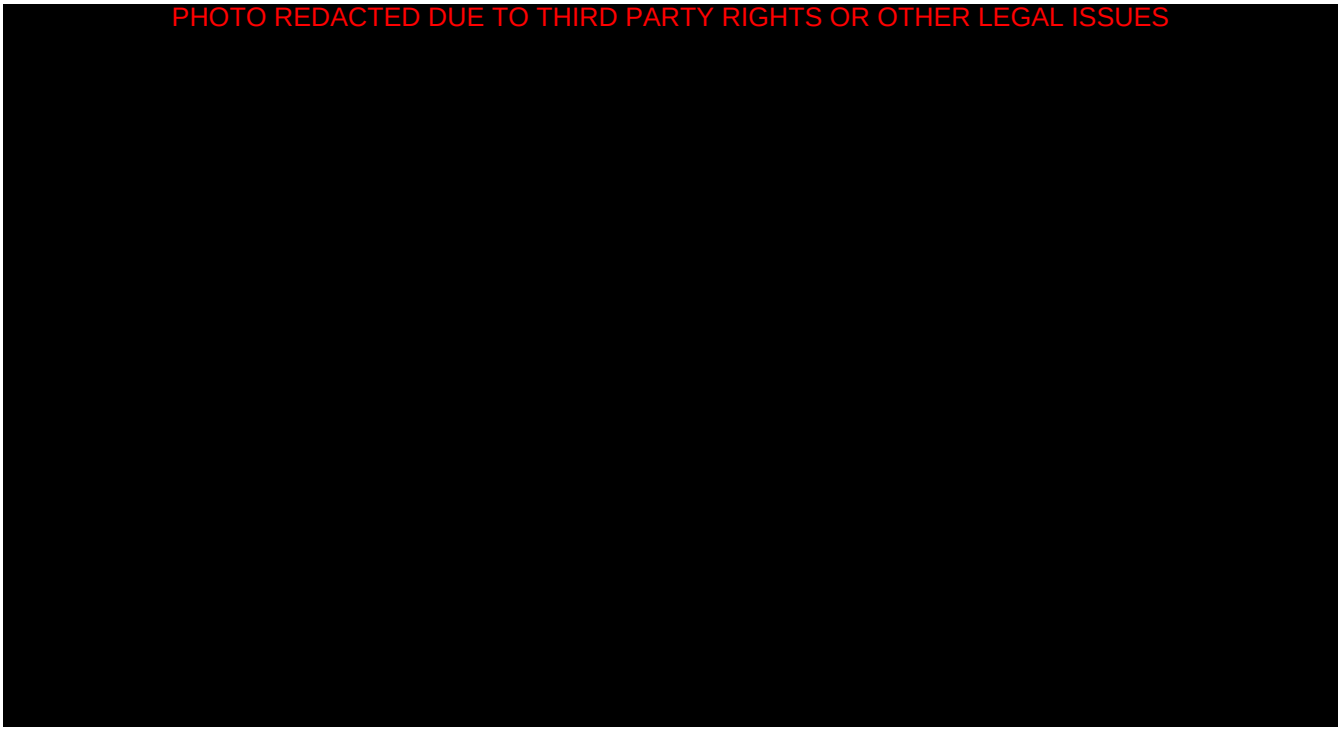
8 Findings from the ‘early adopter’ strand

8.1 The early adopter dissemination event

A dissemination event was held in Nottingham on 21 October 2015[24]. The event consisted of two conference poster sessions, in which the 48 teacher-led RCTs and other forms of experimental research were presented, and oral presentations from:

- Bishop Challoner TSA
- Kyra TSA
- Westbridge TSA

Figure 8.1.1: Teachers who had conducted research sharing their conference posters



During the poster sessions, delegates discussed each other’s projects and findings. The delegates shared ideas on methodologies and areas of teaching that they had investigated. Delegates then voted for the best and most effective presentation of conference poster results. The winning poster and runner-up poster can be found in figure 8.1.2 and figure 8.1.3.

Focus groups were also held in order to capture the schools’ experiences of conducting this type of micro-enquiry and how they felt this research could be of benefit to schools and teachers.

[24] Link to NCTL blog post

Figure 8.1.2: First-place conference poster at the Early Adopter dissemination event

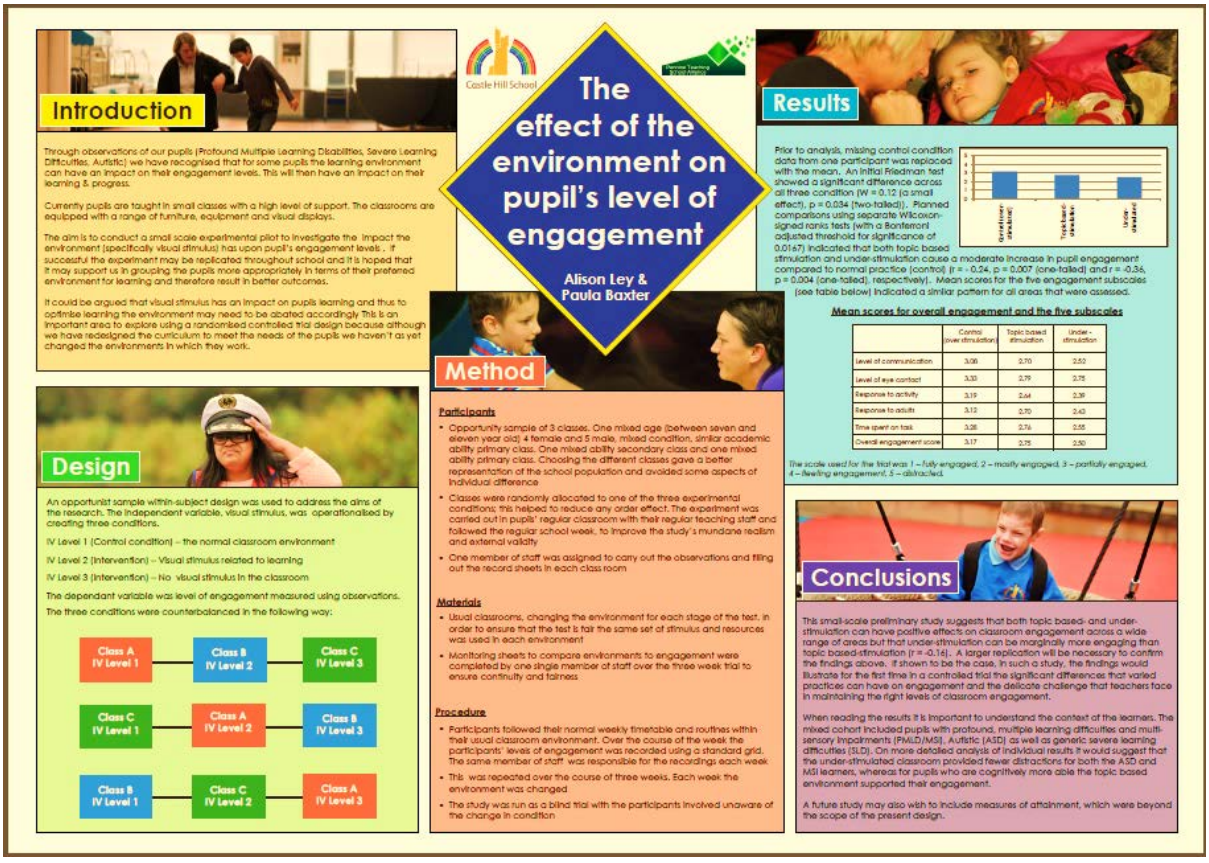
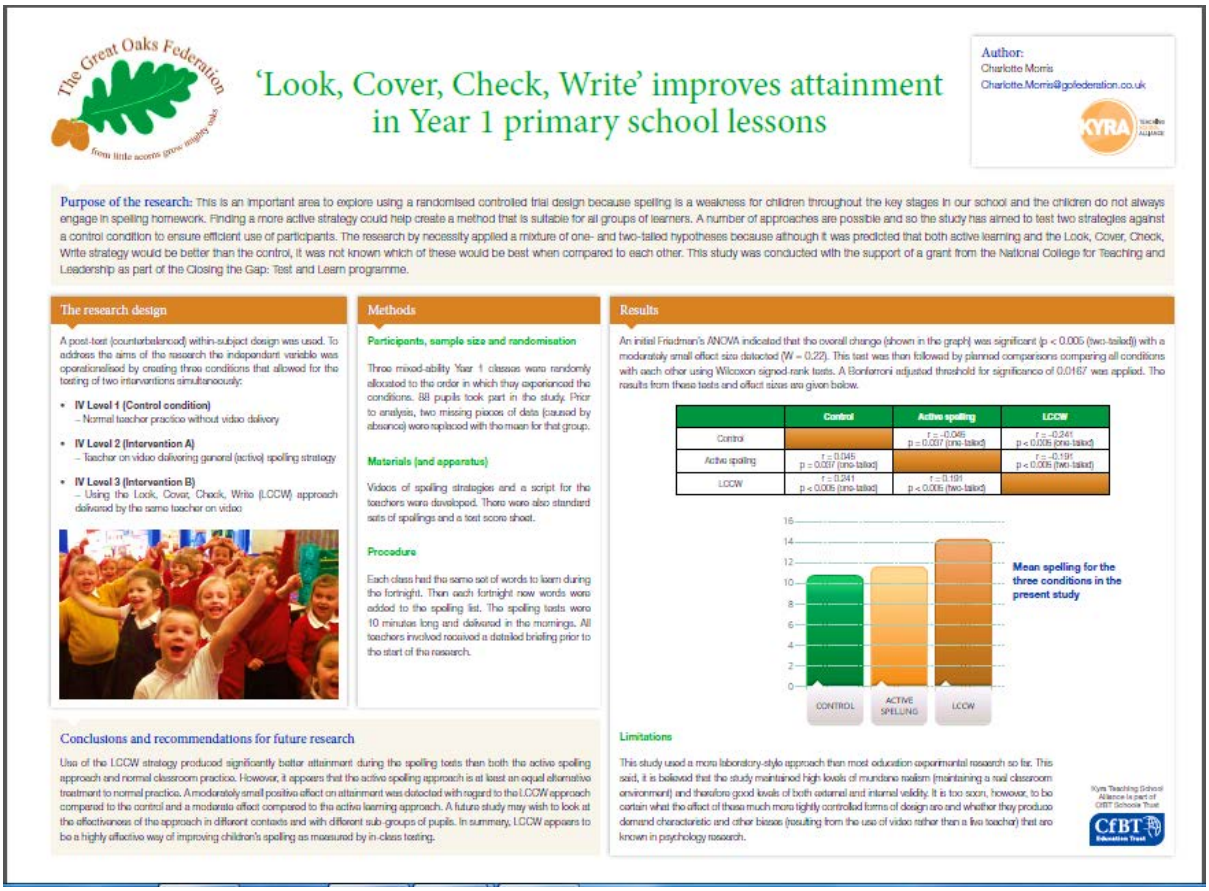


Figure 8.1.3: Second-place conference poster at the Early Adopter dissemination event



8.2 Focus group findings

During the event one-hour focus groups were conducted in groups of approximately 12 delegates. The groups were asked to explore a number of questions with regard to the potential of teacher-led RCTs and their potential to add to the development of an evidence-informed teaching profession. The focus groups' discussions are summarised below.

8.2.1 What have you learnt from being part of the early adopter programme?

Being part of the programme helped participants to develop specific research skills such as data analysis, report writing, controlling variables, impact measurement, proposal writing, research design and planning. It also helped participants develop knowledge in relation to research methodology. Some participants also reported that involvement had made them more reflective, broadened their minds and challenged their perceptions.

Key learning identified by the focus groups included:

- the importance of critically engaging with research to improve practice
- an increased understanding of research processes and the challenges and benefits of conducting research in schools
- recognition of the limitations of research and of data
- the empowering nature of research engagement for teachers and leaders
- the benefits to teachers and learners of embedding a research culture in schools
- that process is a valuable outcome

8.2.2 Has engagement with experimental research changed your perception of evidence-based practice in education (if so, how) and what next steps could you (or will you) take?

For some, engagement in the early adopter programme confirmed pre-existing attitudes towards the importance of evidence-based practice in education; for some it made them realise that small-scale enquiry was valuable and valid if it is well designed; for others it reinvigorated their interest in research and its relevance to the classroom. A small number of participants stated that they had previously been unaware of evidence-based practice and that involvement had 'opened their eyes'. As one participant put it:

> 'I feel it's brought it back down to classroom level for me. I never used to read educational research as it seemed far too removed from what I was seeing on a daily basis – this has changed my perception.'

Other participants spoke about how involvement in the early adopters programme had given them the tools to engage more effectively in research practice or how it would enable them to develop a research culture in their school. A small number of focus group participants spoke about how involvement in the programme had influenced pedagogical debate within their schools.

Some of the next steps participants thought they would take included:

- developing existing studies further – for example through lengthening the timescale of enquiry or increasing sample sizes
- developing greater understanding of the methodological approach of RCTs and micro-enquiry
- identifying new lines of enquiry
- rolling out research and enquiry through their school
- building capacity within and across groups of schools
- seeking funding, or setting aside budget, for more research projects in school.

> ‘Research thinking is what you normally do in the classroom. Actually being engaged in the formalised process has brought into conscious awareness what outstanding teachers do by default.’
>
> Event participant

8.2.3 What do you see as the potential application for teacher-led research of this sort in the future?

Participants recognised the benefits of being involved in teacher-led micro-enquiry and suggested that the approach should be introduced during teacher training, in early career and new leader CPD programmes. They were keen to get more teachers involved and would make excellent advocates for the approach. Participants also noted that it was important to develop ways to easily and systematically share the findings of individual enquiries. The potential power of the approach was summed up by one participant, saying:

> ‘In ten years’ time we won’t be having a conversation about teacher-led research, it will be a defining characteristic of what a school does’

– especially if, as others suggested, it is linked to school development plans.

Focus group participants felt that teacher-led micro-enquiry has a range of future applications, including:

- testing new interventions and policy ideas before implementation

- testing the impact of policy changes imposed from above
- collaborative enquiries across different schools, alliances and contexts
- replicating micro-studies in other contexts
- building an evidence base of what does and what does not work
- driving school improvement and linking to performance management
- as a central tool in CPD

Focus group participants were also asked to share any additional thoughts. There was a considerable appetite to find effective ways to share the research from the programme and, beyond that, to build local and national networks of research-active teachers.

Participants thanked the programme team and valued the time and opportunity to take part and to share their experience with others – in some cases it had triggered bigger conversations about research and evidence for schools and TSAs.

A proposal to strengthen the evidence base arose during the focus groups: to establish a central 'expert' body that school-based researchers could feed the results of small-scale projects into and which, in turn, would then be 'packaged' into projects for other schools to replicate in order to collect more evidence of impact.

Teachers also spoke about some of the problems associated with research in schools. They spoke of the difficulties of finding time to develop and conduct research where there is little or no budget; the difficulties of accessing research materials and support; and of issues around the dissemination of findings, especially in influencing decision-makers to take the findings of research seriously or of convincing others to apply the findings of research to their work.

8.3 Effectiveness of teacher-led randomised controlled trials

Something which makes the early adopter programme unique is that it comprises more than 50 small, individual projects – it has greater value in that the findings are generalisable, because the calibre of the design and the quality of the analysis of the micro-enquiry projects made it possible to draw together findings from each of the projects to directly compare them. In doing so the following was noted:

- 96 per cent of the teacher-led studies yielded a positive effect and only 4 per cent a negative one
- the average effect size, when converted to d for meta-analysis, was d = 0.53 – a medium positive overall effect
- 23 per cent of the studies yielded large positive effect sizes with some very large effects, greater than 1.0

- 55 per cent of the results were significant (despite the relatively small sample sizes). This represents a far greater proportion of significant findings than are generally found in larger-scale more extended RCTs which have taken place over a longer timescale.

The effectiveness of the research designs was almost certainly the result of the teachers using tighter, more controlled designs over a shorter period – designs which are more akin to laboratory psychological studies. In addition, the tests chosen (or developed) by the teachers were generally more closely related to the area of study and the manipulation of content within their experiments. In contrast, some larger-scale trials have deployed measures which are not directly related to the intervention in hand and as such could be termed 'surrogate measures'. A final reason for the nature of the results described above may relate to the nature of the activities the teachers carried out and the interventions that they designed and used. In most cases teachers chose things to test which were directly related to a specific school improvement challenge within their context and developed interventions which (based on the local knowledge and understanding) they believed to be likely to work.

Another indication of the effectiveness of the approach was the variety and creativity of the types of research design that teachers deployed and made use of. Across the 48 studies that were completed in time for presentation at the dissemination event, there were examples of just about all of the commonest forms of experimental research used in fields such as psychology. These included:

- between-subject designs (independent measures)
- within-subject designs (repeated measures)
- matched-pair designs (randomised and non-randomised)
- quasi-experimental studies on already existing groups
- pre- and post-test as well as post-test only
- some designs with three conditions (control versus two different interventions at once)
- some double- and single-blind trials as well as one factorial design.

8.4 Other learning with regard to the development of the researchers as a cohort

Being part of a larger cohort of people taking part in a similar activity has an impact on individuals. The sense of pride and achievement, and the opportunity to see what others have done, sparks further ideas leading to an almost infectious state of enquiry and curiosity – of people saying 'I could try that' or 'you could do this' or 'I can combine that with what I've done and we can test it this way'.

Another of the advantages of the approach, identified by Durham University in the closing remarks, is how enquiry on this scale builds on the deep and specific subject knowledge held by the teachers – this is something that can be lost when we generalise across contexts in larger-scale research studies.

> ‘Probably the most important finding is that there is capacity within the system for teachers to do rigorous, small-scale enquiries that identify whether or not approach A is better than approach B. That is really empowering – if we are going to move towards a school-led system of school improvement, that part of the ecology needs to be there. It still needs larger-scale studies and other kinds of research, but schools definitely need micro-level enquiry to test what they think is going to be effective and make sure it actually is.’

Evaluation from feedback from this strand of delivery includes the following comments:

‘*A huge thank you for organising such an opportunity. This has been so empowering on a personal level but will also impact on the future of education in such a positive way!’*

‘Many thanks for leading us on this journey. It has probably been the biggest success of our alliance so far and led to incredible growth and professional development.’

‘Fascinating to explore the research undertaken by other schools; the impact and implications for practice raised lots of questions to take back and explore further. Highly motivating as a professional. Thank you!’

‘Brill! We will be sad to be carrying on without regular contact with the presenters of this programme. It has been such a privilege to participate.’

8.5 Next steps for teacher-led experimental research

The early adopter programme has clearly shown that there is room within schools to roll out the strategy as a general approach to micro-enquiry, and develop a greater number of teachers’ skills. More work is required to build and develop capacity and there is a need for different variants of micro-enquiry combined with larger-scale enquiry if we are to achieve a stronger teacher-led, school-led system of improvement. Micro-enquiry has a bright future – through micro-enquiry teachers can take control, take responsibility; it empowers them to make more complex and sophisticated professional decisions within their own professional lives.

Several implications emerge from the findings described above. The approach has much potential as a means to develop teachers’ scientific literacy so that they can more easily engage in and develop an understanding of this type of evidence as the number of RCTs increases. This is far from a theoretical point. Indeed, as the recent review by Queens University Belfast shows (referred to in section 2.3), over 800 university-based education RCTs have taken place in the last 10 years. In England alone the EEF has commissioned

over 100. If teachers are going to be able to engage with this research as part of an evidence-informed profession, then building experimental research understanding into teacher training is going to increasingly become a priority, as it is in the medical and healthcare professions.

A second implication and potential of the approach becomes clear when one considers the strong role of context within the education evidence debate. Teacher-led RCTs have much potential as a means of mediating and exploring the effects of prior large-scale studies in the teacher's own context, so that external solutions can be appraised according to local circumstance and priorities for spending. By extension, such approaches could also become a powerful way for schools to pilot changes in pedagogy before they are rolled out across a whole school or TSA.

Finally, not only are teacher-led RCTs desirable professionally, they could also be very cost effective, particularly if teachers can collaborate across schools to build larger sample sizes (paralleling the concept of 'team science' that has grown up within the natural sciences in recent years). For example, one study led by Kyra TSA pooled 11 classes within 10 primary schools, enabling a sample size of 231 pupils. It is not difficult to imagine how, with a small amount of central resources, even larger trials could be constructed in collaboration with teachers and at considerably less cost than many of the large-scale trials that have been delivered in recent years.

9 Conclusions

This section seeks to draw together findings from the project as a whole.

9.1 The programme – a critical academic perspective

This section summarises the discussions in three draft journal articles being produced by OUDE that explore a range of learning from the programme. Conference paper versions of these articles were delivered at the BERA conference in Belfast (15–17 September 2015).

Based on a retrospective analysis of key elements of the programme and its innovative nature, the articles drew upon:

- a range of documents and training materials developed for use within the programme
- questionnaire returns resulting from surveys of the participating teachers
- interviews with a range of the stakeholders and contributors
- feedback and evaluation data from the training sessions

The section firstly discusses findings regarding the policy origins of the programme and motivations for getting involved, before considering methodological issues and the authors' views on how the programme built research capacity in schools.

9.1.1 Policy origins and participants' motivations

The articles highlight how the project was unique in its scale and in its ambition to lead to significant improvements for underachieving students and to support the development of research capacity in the participating teaching schools.

The OUDE team sought to identify the 'policy origins' of the entire scheme, the ways in which it emerged out of the development of teaching schools, the 'closing the gap' objective of the then coalition government and the desire to increase research capacity within the teaching workforce, as well as other elements. By tracking changes that occurred during the period of the project, attention was given to the dynamic nature of an ambitious project of this sort and the ways in which policy is enacted in a major national 'school-led' initiative.

The papers identify the motives of the key stakeholders involved in setting up the project (Durham University, NCTL, CUREE and CfBT) and suggest they remained consistent throughout the project: 1) the opportunity to implement what was described as a series of RCTs (done at scale, collaboratively); 2) the development of research capacity in schools; and 3) the extension of evidence-based practice through teachers' and schools' collaboration in research. The OUDE team shared the last two objectives; perhaps because of their involvement after the initial design phase, they were more interested in

mixed-methods research evidence than the introduction of RCTs as a school research methodology.

Examining the motives for participation reported by TSAs interviewed in this element of the study, and the teachers within them, gives a much more mixed picture. Analysis presented in the papers suggests all were committed to evidence-based practice and to improving understanding of research to close the gap but, even within this sample, there was a wide range of motives and indeed existing practice. For some teachers, either as trial co-ordinators or as teachers from schools allied to a teaching school, the project was viewed as an important part of establishing a working relationship across their alliance. Others, mostly from the early adopters group, came to believe that the real importance of the project was in developing their capacity to implement either a mini-RCT or more mixed-methods approaches to research. In their papers the OUDE team report how, as the scheme developed, there seemed to be a divergence of motives, with classroom teachers who were implementing the interventions valuing the outcomes they could see in their classrooms, and headteachers valuing the results of the RCTs in their decision-making roles.

9.1.2 Methodology

The BERA conference papers also sought to explore the extent to which the overall methodology met the criteria frequently associated with the term ‘randomised controlled trial’ in education. Although this was a very large-scale initiative, the actual interventions were each carried out with relatively small numbers of pupils in a very diverse range of contexts. Policymakers during the recent past have expressed enormous enthusiasm for RCTs, especially in the wake of Ben Goldacre’s paper commissioned by a former secretary of state (Goldacre, 2012) and with the promotion of RCTs by the EEF. In the conference papers the authors suggest some conceptual and technical difficulties in this project which might call some of the RCT precepts into question.

The authors identify two key distinctive features that mark this out as a very innovative project – one that is possibly unique in research within the UK. The first is the scale of the project. They note how there are other major cross-sectional cohort studies, often longitudinal in nature, that enable and facilitate broad sociological and some educational insights to be gained (such as youth cohort studies). But the OUDE team highlight this project, in being based on some specific classroom interventions, as the first large-scale quasi-experimental study that aims to be able to relate particular pedagogical innovations to student outcomes on a large scale.

The second aspect identified by the team as making the project distinctive is the aim to develop research skills and research capacity among the professional staff working in the schools. They recognise that the design of the programme, with its training rounds and networking events, had the overall intention of building an enduring legacy of research disposition and expertise that would enable teachers to continue with their own research

projects into the future, whether within single schools, within alliances or in broader networks. They note that in this respect, the closing the gap: test and learn programme is quite different from much of the early EEF work; however, it is recognised that more recent EEF initiatives such as research use in schools[25], do seek to stimulate just this kind of development. The papers also report on high levels of understanding, both of experimental design in general and the specifics of RCTs.

The OUDE team recognised that many TSAs took the ideas from the RCTs and used these to develop their own experimental small-scale RCT-like research designs. Some 50 of these received further research funding to develop these interventions. These alliances were able to make informed choices about whether to use between-subject or within-subject designs. The NCTL acknowledges the need for a more rigorous understanding of the importance of implementing interventions and tests at the right time as well as ensuring the separation of control and intervention groups to avoid cross-contamination of results. Nevertheless, the fact that many teachers and alliances are involved in collaborative experimental research leads the authors of the research papers to suggest that the project has had a potentially significant impact on the research capacity within schools and, in conjunction with universities as well, on schools' ability to engage with other research from their own contexts and circumstances. In their papers the team conclude that schools with the most advanced understanding of experimental methods also developed the most sophisticated contextual qualitative methods for a closer understanding of the reasons for positive or negative effect sizes.

From a policy point of view they suggest this element is highly significant in England as it places the project firmly within current debates about a reformulation of teachers' professionalism and identity. For example, the proposals for a College of Teaching, the promotion of teachers' research engagement by the Teacher Development Trust (TDT, 2015), the emergence of researchED, the publication of the BERA-RSA report in 2014 calling for research-rich schools and for the development of research literacy among all teachers (BERA-RSA, 2014) – all of these are consistent with a move to develop teachers' research literacy in England and elsewhere. In their research papers the OUDE team also talk about how, within this project, their university-based team has worked in very active partnership with the other team members to bring a particular set of expertise into the mix which has not necessarily been present in some of the other schemes. It is further suggested by the authors that by actively engaging the teachers, rather than seeing them as the 'tools' of an externally devised RCT, the design of this RCT-like project sought to overcome the issue of teachers as consumers of research that may be inherent in RCT approaches to educational research – at least as conceived conventionally. The articles positively express the hope that the extended mixed-methods research capacity developed through the closing the gap: test and learn programme means that evidence-

[25] Link to research use in schools

based practice will, in future, lead to further engagement and involvement in both large-scale and locally contextualised research.

9.1.3 Developing research capacity in schools

In their articles and presentations the OUDE team considered the extent to which evidence emerged from the project suggesting that teachers in schools were becoming increasingly research-literate (BERA-RSA 2014) and that a 'school-led system' could develop research capacity through engagement in a scheme such as this.

They report that data has shed valuable light on ways that the closing the gap: test and learn programme acted as a catalyst on engagement in, and with, research in schools. In their papers they identify some challenges for further development of research-informed schools. First, whilst the complexity of TSA structures and links provided the economies of scale to support research leads, it also suggested a degree of fragility in the process. Different organisations and practices can have different motives and agendas, which, in turn, may make research difficult to organise and support – especially without the guidance offered through the project.

Whilst there are a range of organisations promoting and supporting experimental-style research, it would be up to the schools and alliances to approach them if needed. The authors also discuss what they consider to be the pragmatic positioning of research with school improvement. They recognise that this provides a strong justification for research within school practices, but express concern that it may dictate both what and how it is researched, alongside or within other aspects of schools' educational aims. It is argued that the organisational epistemology of research is located within wider organisational values, and current perceptions of research rigour may not fit easily alongside potentially conflicting agendas.

The research papers continue by examining the wider cultural and political issues as to how teaching is conceived, as well as how research is conceptualised across schools and alliances, and across wider links with other organisations. The authors suggest this is about the conflicting perceptions of teaching as much as it is about the place of research in schools and the involvement of other organisations within this. At one level, they argue, this is only to be expected, as many areas of research have their own particular concatenations of methodology and institutional organisation – for example, the involvement of universities and pharmaceutical companies in medical research; however, the particular issues raised by the authors here are compounded by wider policy shifts in school structures, higher education and the research economy. In the long term, school-led RCTs may flourish and develop as a very significant form of educational research, although the analysis presented by the OUDE team would suggest that different schools and alliances will position it differently, and it will become part of a 'buffet' of research methods.

9.2 The national dissemination event focus group findings

A national dissemination event was held in London on 18 November 2015. The day had two aims:

- to celebrate progress and share learning and impact, good practice, successes and challenges
- to develop next steps for practitioner research/school-led trials.

During the day two 30-minute focus sessions were led by facilitators from OUDE. The first session focused on exploring the benefits and successes of the programme. It also included a discussion of next steps and how approaches like those used in the closing the gap: test and learn programme might be taken forward in the future. The second session sought to elicit participants' views regarding issues that had arisen during the programme, barriers to implementation and the sort of solutions that had been adopted. The focus groups included group discussion and the ability of participants to log their responses online during the sessions so that more in-depth scrutiny of the evidence could be made later. The focus group discussions are summarised below.

9.2.1 Benefits

Participants reported a number of benefits for their TSAs. A significant benefit of involvement in a project of this sort was the heightened awareness of research methods, of the training available and of the research capacity in their own and neighbouring alliances. They recognised that there were now schools that have done research, and had had success, on which they could call for support. Some participants also spoke of how they now had a 'shared language' around research which improves understanding and communication of ideas. Linked to this was a shift in teachers' views of what research is – with some participants suggesting that their views had 'radically' changed – as participation in the project had demystified research.

Importantly, a number of participants spoke of the impact on learning and on teacher behaviours. For example, it was suggested that involvement had had an impact on teaching assistants' and teachers' subject knowledge, that it had led to a better understanding of how children learn and that, through the use of standardised diagnostic tests, it had had a positive impact on assessment and testing. Others spoke of how involvement in the project had made an impact on teachers' perceptions of themselves, their work and their role in their own professional development. Finally, an important benefit in an age of self-improvement was an increase in collaboration within schools, across and outside of alliances.

9.2.2 Successes

Successes identified in the feedback sessions predominantly focused on pupil outcomes. For example, it was reported that involvement in the project had resulted in:

- improved attainment
- increases in pupils' self-confidence
- increases in pupils' motivation
- the empowerment of pupils – giving them more control to direct their own learning and giving them a new vocabulary to talk about their work and learning.

Involvement in the interventions had also given schools experience and subsequently the confidence to embed strategies across the school (e.g. literacy across the curriculum). Another reported success relates to enhancing teaching schools' core responsibilities for ITT and CPD with the interventions developed through the project being integrated into the schools' practice.

9.2.3 Next steps

The project has generated enormous enthusiasm and interest in research and some suggested that this might be maintained through research groups in schools or possibly by identifying 'research champions' in departments or schools. Participants suggested that what is now needed is a commitment to 'scaling up' – both in relation to the interventions and in relation to research activity in general. Participants also suggested that this will require greater 'scaffolding' and more funding.

With regard to the interventions themselves, participants suggested that there is a need to recognise wider impact, beyond what can be measured by standardised tests as well as a need to assess impact over longer time periods. There was also a desire to make the whole approach more inclusive so that all students would ultimately benefit.

Some participants had suggestions for next steps linked to professional development, for example, how teachers might gain accreditation (towards higher degrees) or for involvement in research activities. There were also suggestions that all CPD for teachers should be evidence-based (or research-based) with some believing it would be valuable to link with performance management.

9.2.4 Issues and solutions

In the second sessions, the participants were asked to consider significant issues that had arisen during the project and how they had overcome them. This session also explored suggestions of how to improve any similar future projects. Issues and solutions fell broadly into three categories:

- those relating to participation by schools and alliances
- those relating to the research interventions
- those relating to the research project overall.

Issues and solutions relating to schools' and alliances' participation reflected the realities of daily school life – participants talked of difficulties getting colleagues on board, staff turnover and schools being inspected during the project resulting in their grading changing; and this in turn leading to a change in senior leadership team focus. Co-ordination of the project across alliances was also an issue raised by some groups. Suggested solutions included: pestering, chivvying and bossing colleagues (nicely), involving other staff (e.g. IT leads and special educational needs co-ordinators (SENCOs), and developing networks across the alliance) and taking a proactive approach to keeping the original enthusiasm going.

In terms of the research process itself, participants raised questions about the design of the large-scale trial interventions. It was suggested that the interventions were not always appropriate to schools' or alliances' needs; that testing instruments were not always age-appropriate or compatible with available IT, and that there had possibly been insufficient time between intervention and test. Solutions that had been developed to overcome some of these issues included teachers inputting data manually, contextualising the interventions and bussing pupils between schools to access computers.

Issues related to the research project concerned the overall design and the available funding. It was felt that the project lacked clarity at the start and the timescales were not helpful – especially given that schools and alliances were at differing points on their research journey. This is discussed further in the section 9.2.5 below. Issues regarding the allocation of funding were also raised, especially with regard to factoring in regional differences (for example, the longer and more challenging travel distances for some regions, such as the south west).

9.2.5 Advice to the National College for Teaching and Leadership

Project participants made a number of suggestions for change, if a similar project were to run again. Chief amongst these was the need for more time between project launch and the start of any intervention – this would allow for more effective planning and preparation and improved co-ordination. It would also allow time for redesign of intervention materials if needed and for teachers to become fully familiar with the approach. A revised timescale may also improve project organisation and enable greater clarity in the early stages of the project; it may also support greater differentiation, recognising that different TSAs and their schools are likely to be at different starting points regarding research in schools. This should be built into project design and could lead to more specific guidelines for participants. It was also suggested that schools would like greater involvement in the choice of interventions, or at least to see a more collaborative approach to the

development of interventions so that they would more closely match the particular needs of schools.

Finally, some suggestions regarding the project design were made. It was suggested that the approach to testing may need to be more sensitive to individual interventions. It was thought to be necessary to decide whether to take a national or regional approach to future projects.

9.3 Summary of programme findings from all the areas of delivery and engagement (including qualitative data from teacher surveys during the life of the programme)

Pooling the evidence from the trials, OUDE research, focus groups and engagement activities, a number of overall conclusions can be drawn:

- the programme has clearly demonstrated the capacity of schools to engage in research through large-scale multi-arm trials and micro-RCTs; this also increased engagement with research and discussion of research findings
- contrary to assumptions made at the start of the programme, schools were not resistant to the use of control groups, engagement in statistical research or to the use of RCTs in general
- teachers can take a more active role in the delivery of RCTs. However, this form of approach requires investment in training and careful control of the communication structure and engagement protocols to ensure that individual trial sites do not become too distant from any middle-tier process used to build scale
- teachers can design and implement teacher-led RCTs focused on local school improvement questions and the evaluation of interventions to determine their effectiveness in individual contexts, paralleling approaches in healthcare
- surprisingly, despite their relatively small scale, teacher-led RCTs frequently produce statistically significant findings – almost certainly the result of shorter treatment windows, tighter controls and the ability to use research designs such as within-subject and matched-pair designs – designs which inherently increase the power to detect an effect
- the teaching profession already has a deep and available resource of individuals with experience of quantitative methods and their application outside of education research (such as science teachers and psychology teachers). The challenge going forward will be to connect these people with one another and provide them with models of how such approaches can be taken forward in school-led education research projects
- the results can be seen as emphasising the challenge of diagnosis within education – in that it is hard to find general things which improve outcomes across the board

(except for NIP). Schools therefore need to think carefully about how their adoption and embedding of research-based approaches will stand more chance of success ('be above average'). Effective use of large-scale evidence is likely to require good diagnosis and targeting – and training of school leaders to understand how to do this

- micro-trials build capacity for research knowledge and engagement with research (the content of research findings) as well as engagement in research (trials design and implementation) which will in turn build capacity for the critique of the robustness of research findings. The evidence from this programme suggests that further development would be a valuable policy proposition

9.4 Policy implications and recommendations

Drawing on the above, the following policy implications and recommendations can be made:

- collaborative large-scale trial approaches such as the one applied in the closing the gap: test and learn programme have much potential with regard to reducing the costs of large-scale trials and thus the evidence base. Such trials in the future should, however, implement the learning from this programme with regard to those areas which need tighter national administration to avoid internal validity issues (for example, data labelling (if schools are given responsibility for this) and consistency in relation to intervention delivery)
- 'toolkit' summaries of research (such as meta-analyses of effect size) are at a broad level effective to some degree, but these need to be tested at a micro level. Such an approach, if co-ordinated nationally, could build an evidence base to support implementation and understand variation related to context
- finding no effect increases the range of choices that can be made, as this suggests that the impact on learning outcomes is the same. This point needs to be emphasised in the development of scientific literacy as policy develops in this area
- as the volume of trial evidence increases, the need to develop teachers' scientific literacy and understanding of experimental research methods, in order to interpret those findings accurately and appropriately, is becoming increasingly clear
- there is potential for a micro-enquiry approach in ITT, supported by good research design and analysis skills, as this would also address trainees' scientific skills about enquiry and statistical capability. With regard to the use of inferential tests, effect sizes and confidence intervals, such training could also improve the quality and efficacy of local school improvement processes

10 Glossary

Term	Definition
ANCOVA	Analysis of covariance.
ANOVA	Analysis of variance.
Between-subject (independent measures)	Research designs in which different people are in the control group and intervention group. Such designs contrast with within-subject (repeated measures designs) where all participants do both the intervention and the control.
Blocked randomisation	Also known as stratified randomisation, this is an approach that attempts to control for between-subject variation by 'blocking' participants to balance characteristics during randomisation. In the case of CtG randomisation was at school level.
Confounding effects	Factors (or 'confounding variables') outside the intervention which may have had such an effect on the results that the trial results represent this factor rather than the difference between the intervention and the control.
Early adopters	Short for 'early adopters of experimental research as a school improvement approach', this refers to the 50 successful grant applicants who designed such research approaches using the materials that were trained during the RDNEs.
Effect size	Within CtG this refers to the amount of change that has taken place when the intervention is compared to the control. The form of effect size used to explain this with regard to the year 1 results is a standardised mean difference known as 'Cohen's d'.
Large effect size	d = 0.80, suggesting that 79 per cent of children in the control would be below average in the intervention.

Term	Definition
Moderate effect size	d = 0.50, suggesting that 69 per cent of children in the control would be below average in the intervention.
Negative effect size	The children's progress as a result of the intervention has been worse than the control.
Positive effect size	The children's progress as a result of the intervention has been better than the control.
Small effect size	d = 0.20, suggesting that 58 per cent of children in the control would be below average in the intervention.
Gain scores	In the analysis in year 1, because of the distributions and pre-test results, gain scores were used to assess pupil progress. A gain score is the post-test score minus the pre-test score calculated at individual pupil level.
Hypothesis	A clearly defined statistically testable statement with two components (a null hypothesis and an experimental hypothesis). The null hypothesis is accepted if a non-significant result is obtained, the experimental one if a significant result is found. For example: Null hypothesis – five months of exposure to lessons using 1stClass@Number does not improve pupil progress as measured by the GLA Progress in Maths Standard Age Score. Experimental hypothesis – five months of exposure to lessons using 1stClass@Number does improve pupil progress as measured by the GLA Progress in Maths Standard Age Score.
Intervention/treatment	In CtG this means the particular training programme that is being trialled against the control.
Outlier	A very extreme score that risks affecting the final results. In the case of the CtG analysis these were removed prior to the analysis taking place.

Term	Definition
Randomised controlled trial	A form of experimental research in which conditions are assigned by the researcher (one of which is considered to be a control condition), and in which participants are randomly allocated to the condition(s) that they will be exposed to, prior to treatment.
Replication	The repetition of a study in order to check whether the results are likely to generalise, considered an essential component in scientific research methods.
Significance	There are a number of ways to define significance. The detail of which is dependent on the depth of statistical definition a writer feels it necessary to include. Brace, Kemp and Snelgar[26] offer a straightforward interpretation, as 'the level of probability (p) that the results are due to chance, at which we reject the null hypothesis and accept the experimental hypothesis'. The null hypothesis being a theoretical statement that there is no difference between the things we are testing and the experimental hypothesis the opposite (i.e. there is a difference). It is, however, important to realise that the significance level set for rejection or acceptance of the null hypothesis (known as alpha) should not be equated with the actual size of any experimental effect.[27] More technically, the American Psychological Association online glossary[28] defines 'a significant difference' as 'a difference between experimental groups or conditions that would have occurred by chance less than an accepted criterion; in psychology, the criterion most often used is a probability of less than 5 times out of 100, or $p < 0.05$.'

[26] Brace, N., Kemp, R. and Snelgar, R. *SPSS for Psychologists* (third edition), Basingstoke: Palgrave Macmillan (2006): 382.
[27] Dancey, C.P. and Reidy, J. *Statistics without maths for Psychology* (fourth edition), London: Pearson (2007): 144.
[28] Research action glossary

Term	Definition
Small-scale trial	A trial that, although it may be able to detect a positive or negative effect size, is of a size that is unlikely to reach statistical significance. For example, for a 0.4 effect size to be significant in a study with two groups, a sample size of above 78 in each group would be necessary.
Sub-group	A specific group of individuals within the particular trial (e.g. FSM children).
Treatment fidelity	The extent to which the intervention was delivered consistently.
Treatment window/period	The period between pre-test and post-test.
Trial	The process of assessing the effectiveness of an intervention against a control.
Within-subject design	Within-subject (repeated measures designs) where all participants do both the intervention and the control.

Appendix A: The completed Early Adopter studies

School	Trial
Aspirer Teaching School Alliance	The scripted intervention programme, Talk Boost, may be significantly more effective at developing the expressive language skills of children in reception than practitioner planned regular small talk groups
The Blue Coat School, The Northern Alliance	Preliminary evidence for the effect of a Mental Toughness intervention programme for year 10 students
Gatley Teaching School Aliance	The impact of supported familiarisation in the transition between phases
Aspirer Teaching School Alliance	A small group intervention, Pulling It Together may be more effective at developing a child's phonic and word reading skills than small group additional 'phonics' lessons
Gatley Teaching School Aliance	A preliminary pilot study indicating a positive effect for the use of tablets in the improvement of phonological attainment
West Essex Teaching School Alliance	Using the visualisation technique of bar modelling does not improve the choosing of an appropriate method to solve word problems in mathematics
Swanshurst School, Bishop Challoner TSA	Peer feedback is equally effective in improving pupil progress in essay-writing at A-level as teacher feedback: preliminary evidence
Gatley Teaching School Aliance	Increasing involvement of teaching assistants in reviewing intervention programmes accelerates progress with SEN pupils to close the gap in mathematics
Redhill Teaching School Alliance	There is little difference in the amount of time students spend revising if they are given time to produce their own revision materials during the lesson
Pennine TSA	The effect of the environment on pupil's level of engagement
Aspirer Teaching School Alliance	Preliminary evidence from a small scale pilot study into the effectiveness of playing games as a means of developing fluency in the automatic recall of times tables

School	Trial
Gatley Teaching School Aliance	Increased accountability through coaching and mentoring, has a positive impact on the progress towards more effective marking and feedback
EOS TSA	The use of kinaesthetic strategies improves spelling in year 2
Worsborough Common Primary School, Barnsley TSA	Directive versus inductive approaches to teaching spelling: Which is the most effective in supporting effective learning and progress in spelling?
Kyra TSA	The impact of domestic help to facilitate home reading
Monks Abbey Primary School, Kyra TSA	Peer reading improves the reading age of pupil premium children compared to reading only to adults
The Heath Teaching and Learning Alliance	Self-selected on line gaming stimulus improves boys' creative writing: the impact of using computer games to promote creativity in boys' writing to help bridge the attainment and gender gap
Westbridge TSA	A preliminary study into the effects of a weekly spelling test on pupils progress in retaining spellings
The Blue Coat School, The Northern Alliance	Preliminary evidence for the effect of increased motivation and competition intervention in mathematics which appears to be equal to existing practice
The Blue Coat School, The Northern Alliance	Preliminary evidence for the impact of context-based learning on effort and achievement within extended writing.
Long Sutton County Primary School, Kyra TSA	Changing a classroom's teaching environment can raise attainment in English and mathematics
Bentley Wood High School for Girls, Harrow Collegiate TSA	The impact of Growth Mindset interventions on students in different key stages
The Great Oaks Federation, Kyra TSA	'Look, Cover, Check, Write' improves attainment in year 1 primary school lessons
Lambeth TSA	Sting in the Tale: a collaborative opera by The Wyvern Federation and English Touring Opera

School	Trial
Cranbourne Business and Enterprise College	Pupils' progress does not increase if effort related written feedback is used in addition to other regular feedback types
The Blue Coat School, The Northern Alliance	Preliminary evidence for the effect of mental toughness intervention to narrow the gap between pupil premium and non-pupil premium students
Weatherhead High School, The Learning, Teaching and Leadership Alliance	The use of tablets and applications in GCSE Spanish raises attainment in listening and reading examinations.
Bentley Wood High School for Girls, Harrow Collegiate TSA	Co-teaching - an investigation into how effectively co-teaching closes the achievement gap for underachieving pupil premium students
West Coast TSA	Evidence from a preliminary non-randomised feasibility study into the effect of Building Learning Power on maths attainment.
The Queen Katherine TSA	Preliminary study into the effects of giving dedicated improvement and reflection time (DIRT) after feedback on written work has been provided
Lambeth TSA	Number skills video project supporting the Oval Cluster number master programme
Ridgeway School, The White Horse Federation	Preliminary evidence from a small-scale pilot study regarding the use of an neuro-linguistic programming (NLP) informed coaching programme to support year 10 English teaching
St Margaret's CE School, Kyra TSA	Verbal and visual-digital feedback on creative writing in rural primary schools improves progress rates compared to written feedback
The Queen Katherine TSA	Learning from a test that we failed to complete
The Vale Primary School, School Partnership Trust TSA	Preliminary evidence regarding the effect of personal interest-based learning on progress in early years classrooms.

School	Trial
The Bishop's Stortford High School, Catalyst TSA	There will be a significant difference in the attainment of students when taught using a flipped classroom method in comparison to those given a traditional lecture style lesson.
Sacred Heart Catholic Primary School & Blue Sky Teaching School Alliance	The effect of physical activity on academic performance of pupils in maths
The Queen Katherine TSA	The effect of mindset training on low and high ability learners – preliminary evidence from a case-matched quasi-experimental study
Gatley TSA	A preliminary pilot study indicating a positive effect for the use of iPads in the improvement of phonological attainment.
St. John's C of E Primary School, West Essex TSA	Aiming for speed: will learning to play darts help to increase writing speed?
Stamford Welland Academy, Cambridge Teaching Schools Network	Flipped learning has a positive impact on attainment in modern foreign languages
Stamford Welland Academy, Cambridge Teaching Schools Network	Using an internet based homework calendar that tracks submissions encourage a higher rate of homework completion
South Bromsgrove High, South Bromsgrove Alliance	A preliminary pilot study into the impact upon student engagement and attainment when directly responding to teacher's feedback
East Kent Learning Alliance	A small scale pilot study to establish whether the use of digital games based technology aids motivation, engagement and attitude to learning in the classroom with a year 9 mixed ability class.

School	Trial
Palmerston School, Palmerston Inclusive Alliance Support	A small scale pilot study into the effectiveness of two meditative techniques on the improvement on concentration in a severe learning difficulties (SLD) school. Does practising mindfulness techniques at the beginning of lessons increase engagement of pupils with SLD?
Blue Sky TSA	The effect of specialist art teaching on improving handwriting
West Coast TSA	Evidence from a preliminary non-randomised feasibility study into the effect of the focused development of learning behaviour on maths attainment

National College for Teaching & Leadership

Reference: DFE-RR500b

ISBN: 978-1-78105-557-1

Any enquiries regarding this publication should be sent to us at: teaching.schools@education.gsi.gov.uk or www.gov.uk/nctl

This document is available for download at www.gov.uk/government/publications