Research Report DFE-RR091a

Every Child Counts: the independent evaluation Technical report

Torgerson, C.J., Wiggins, A., Torgerson, D.J., Ainsworth, H., Barmby, P., Hewitt, C., Jones, K., Hendry, V., Askew, M., Bland, M., Coe, R., Higgins, S., Hodgen, J., Hulme, C., Tymms, P.

Department for Education

Acknowledgements:

We acknowledge the following:

Funder: Department for Education (DfE).

Project Manager: Jenny Buckland (DfE), Karin Bosveld (DfE,), Caroline Halls (DfE), Alison Pollard (DfE).

Steering Group Membership: Jenny Buckland (DfE), Nigel Bufton (National Strategies), Ann Dowker (University of Oxford), Nick Dowrick (Edge Hill University),Susan Fletcher (DfE) Jean Gross (KPMG), Di Hatchet (Every Child a Chance Trust), Alison Pollard (DfE), Amanda Sasia (DfE).

Research Advisory Group Membership: Peter Tymms (Durham University), Carol Aubrey (University of Warwick), Margaret Brown (Kings College London), Mark Newman (Institute of Education), Robert Slavin (University of York), Maggie Snowling (University of York).

University of York: Ben Cross (randomization programme), Arthur Kang'ome (statistical support), Louise Elliot (data management).

Durham University: Kate Bailey (Pair site visits)

Programme Development and Professional Development (Numbers Count): Edge Hill University and Lancashire County Council; Nick Dowrick (Head of Every Child Counts Programme); ECC National Trainers; ECC Teacher Leaders.

Group work Training: Marie Heinst and Mary Clark, Every Child a Chance Trust Mathematics Consultants.

All independent testers, schools, headteachers, Numbers Count Teachers, class teachers, teaching assistants, pupils and parents/carers who took part in the evaluation; support staff at the University of York who entered the data and helped with administration.

Conflict of interest statement:

The authors have no known conflicts of interest. This is an independent evaluation.

Technical Report

Contents

Tables

Figures

Glossary

CC: case control

CCD: case control design

CEA: cost-effectiveness analysis

CI: confidence interval

CONSORT: Consolidated Standards on the Reporting of Randomized Trials

DCSF: Department for Children Schools and Families

DfE: Department for Education

DfES: Department for Education and Skills

ECaR: Every Child a Reader

ECC: Every Child Counts

ES: effect size

FSM: Free School Meals

FSP: Foundation Stage Profile

HSSEC: Humanities and Social Sciences Ethics Committee

ICER: incremental cost-effectiveness ratio

IDACI: Income Deprivation Affecting Children Index

INC: individual Numbers Count

ITS: interrupted time series

ITT: intention to treat/teach

KS1: key stage 1

LA: local authority

NC: Numbers Count

NCT: Numbers Count teacher

NPD: National Pupil Database

OLS: ordinary least squares

PIM 6: Progress in Mathematics 6

PIPS: Performance Indicators in Primary Schools

PNC: pairs Numbers Count

QED: quasi experimental design

RCT: randomized controlled trial

RR: Reading Recovery

SD: standard deviation

SDQ: Strengths and Difficulties Questionnaire

SE: standard error

SEN: special educational needs

SRA: Social Research Association

TA: teaching assistant

TIMSS: Trends in International Mathematics and Science Study

TL: teacher leader

TNC: triplets Numbers Count

UT: usual teaching

Y2: Year 2

Chapter 1: Background and context

1.1: Introduction

The relative improvement of primary mathematics teaching has been widely accepted, with the number of 11 year-olds gaining level 4 and above at key stage 2 having risen from 59% in 1998 to the current figure of over 77% (Williams, 2008). However, the picture for low achieving pupils is of widespread concern. Since 1998 the number of children failing to achieve level 3 has remained at about 6%. Whilst the majority of children have improved, the lowest performing children have remained at much the same level (Williams, 2008).

There are many harmful consequences for pupils of low attainment in mathematics: in the short term (for example, having a negative effect on learning in a variety of areas of the curriculum, including mathematics itself), and in the longer term (for example, difficulties at secondary school and into adulthood). Slightly higher than 5% of lower attaining pupils at key stage 1 (KS1) go on to leave secondary education with no qualification at all in mathematics.

It is widely agreed that a child who is having significant difficulties at an early stage (i.e., during KS1) is likely to under-achieve in mathematics throughout their school life, and beyond. To help address this problem, the Primary National Strategy (PNS) introduced the three wave model of intervention in mathematics, with provision for the lowest performing children to receive personalised, individual teaching.

It is against this backdrop that the KPMG Foundation initiated the formation of the Every Child a Chance Trust in 2007. The trust seeks to provide a solution to the problem of underperformance in mathematics by the lowest attaining group of young children. The primary aim of the trust is to provide benefits to the children themselves and to the long term economic performance of the country as a whole.

In 2007 the Every Child Counts (ECC) partnership was formed, comprising of the Every Child a Chance Trust, The National Strategies and the Department for Children, Schools and Families (DCSF) (now the Department for Education, DfE), and was joined by Edge Hill University in 2008. Later in that year (2007) the then Education Secretary, Ed Balls, announced[1] a total of £144 million funding for both ECC and the sister initiative in reading - Every Child a Reader (ECaR). This committed the government to providing ECC (and ECaR) on a one-to-one basis to 30,000 six year old pupils by 2010/2011.

The main aim of ECC is to develop and support an intervention for the lowest achieving 5% of KS1 children, with a subsidiary aim of impacting on standards more widely by influencing classroom practice and supporting less intensive (teaching assistant led) interventions for the low achieving 5-10% group. Much of the underlying pedagogical rationale of ECC was informed by a DfE-sponsored report 'What Works for Children with Mathematical Difficulties?' (Dowker, 2004). This report helped to bring together the evidence base for effective interventions. The conclusions of the report were that children who are underperforming in mathematics are highly susceptible to targeted intervention, and that

[1] www.dcsf.gov.uk/pns/DisplayPN.cgi?pn_id=2007_0171

intervention should take place at an early age to reduce negative attitudes and allow access to other aspects of the curriculum.

The ECC initiative develops mathematics interventions for Year 2 children within the following three waves:

Wave 1 - Quality classroom teaching for all children;

Wave 2 – Small group additional intervention for children just below national expectations;

Wave 3 – Individual or very small group intervention with a trained and supported TA for children who are struggling, and additional intervention on an individual and/or very small group basis with a trained specialist teacher.

The key ECC programme, Numbers Count (NC) provides an intensive one-to-one intervention for those children identified as lowest achievers in Wave 3 (see above). In practice it aims to raise their level of performance so that they achieve level 2c or higher, and wherever possible level 2b or higher by the end of KS1 – in effect putting them on a par with their peers, and enabling them to continue to progress in mathematics in the normal mainstream class setting.

The ECC programme contributes funding to help schools to employ and train specialist Numbers Count teachers (NCTs) to deliver daily one-to-one intensive NC teaching for those children in the lowest 5% with the greatest difficulties.

Alongside the early development of ECC, the Review of Mathematics Teaching in Early Years Settings and Primary Schools led by Sir Peter Williams commenced in 2007 and reported the following year. It largely sought to build on the Primary Framework for Mathematics and the Early Years Foundation Stage, and one of the six key areas considered was a review of ECC:

> "The review should specifically make recommendations to inform the development of an early intervention programme for children (aged five to seven) who are failing to master the basics of numeracy – **Every Child Counts** - as recently announced by the Prime Minister." (p.2)

The review team was able to observe the ECC initiative at its research stage, and it should be noted that the intervention itself had not yet been formally identified / finalised at that time. Their recommendations did, however, go on to inform the development of Numbers Count[2].

1.2: Numbers Count

Edge Hill University, working in partnership with Lancashire County Council, developed Numbers Count (NC).

NC is a 12 week programme, consisting of daily 30 minute one-to-one sessions for the target children, delivered by specially trained Numbers Count teachers. The core elements are a comprehensive diagnostic assessment of each child's strengths and weaknesses, core learning objectives for the lessons and guidance for teachers on lesson structure and key teaching approaches. NCTs are supported by a continuing professional development

[2] Numbers Count is a trademark of Edge Hill University.

programme and a quality assurance system. NC is specifically designed to help children to develop their knowledge and understanding of number. In addition, NCTs aim to give children confidence in number and an understanding of patterns and relationships so that they can extend learning to other aspects of mathematics in their class lessons. They use shape, space and measures, and handling data as contexts for the development and application of children's number skills, and children continue to study the full breadth of the mathematics curriculum with their class teacher. (For more detailed information about Numbers Count please see the Process evaluation Appendices, Appendix 2, Torgerson et al, 2011c).

1.3: The evaluation

There are two main strands to the evaluation: an impact evaluation and a process evaluation. The impact evaluation used experimental and quasi-experimental methodology to establish effectiveness. Two randomized controlled trials (Trials 1 and 2) evaluated, firstly, the effectiveness of receiving NC compared with not receiving NC, and, secondly, the relative effectiveness of NC compared to an adapted intervention based on NC delivered to pairs or triplets of children. The trials were undertaken in 65 schools across the country. A series of secondary analyses used interrupted time series and case control designs to support the results of the two randomized controlled trials by examining the impact of ECC more widely. The process evaluation used a cross-sectional design to support the trials by seeking to explain, better understand and contextualise their findings, as well as considering wider pedagogic and organisational (both school and programme level) issues.

1.3.1: Trials 1 and 2

There is a clear need to obtain reliable evidence to inform policy and practice, and crucially to establish the level of effectiveness of NC compared with normal classroom practice. The best way of obtaining reliable evidence for the effectiveness is through the use of a randomized controlled trial (RCT) design. Other potential designs for evaluating NC, such as a single group pre and post-test design would be confounded through temporal changes (the natural process of children improving their mathematical skills through ordinary teaching and or increasing maturity) and regression to the mean effects (the statistical phenomenon where children who are tested and achieve scores at the extreme of a distribution will, on average, tend to show an improvement on re-testing irrespective of any real change whatsoever). It is widely acknowledged that single group pre and post-test studies exaggerate estimates of effectiveness in the order of 60% or more when compared with studies that include a contemporaneous control group (Lipsey and Wilson, 1993).

As well as having a contemporaneous control group it is also crucial that such a control group should be assembled through the process of random assignment, otherwise bias can be introduced. Such bias can either underestimate or overestimate the effectiveness of an intervention. For instance, if pupils who receive NC were selected on the basis of a low score on a test then regression to the mean effects will ensure an exaggerated improvement compared with children in the control group who scored higher on the pre-test and do not receive the intervention. Random allocation ensures that such biases are absent from effect size estimates.

Finally, it is also important that tests undertaken post randomization are administered by personnel who are *blinded* or *masked* to the membership of the intervention and control groups. This is to avoid conscious or unconscious effects by the testers who may have a desire to ensure that the intervention children perform to the best of their ability and consequently the results may not be a fair reflection of the performance of intervention and control children. The two trials described in this report have been designed to be rigorous randomized controlled trials. We used the CONSORT guidelines (Moher et al 2001) to minimise all potential biases through their design, conduct and reporting.

1.3.2: Secondary analyses

The secondary analyses involve a comparison phase using national data and two quasi-experimental designs: interrupted time series design and case control design.

In an interrupted time series (ITS) design a group of participants is tested repeatedly both before and after the introduction of an intervention, in this case before and after the introduction of ECC in two cohorts of schools. In essence this is a single-group, pre and post-test design with multiple before and after measurements which enable confounding variables (regression to the mean effects, temporal changes) to be detected. If the plot of the dependent variable (KS1 outcomes) shows a change in level or direction at the point of intervention (immediately after or shortly delayed), and potentially confounding variables have been minimised due to multiple observations (multiple schools), then it is possible to ascribe a causal relationship between the intervention and the dependent variable (KS1 outcomes). However, it should be noted that the ITS design does not permit such a strong causal relationship to be established as the more rigorous RCT design. For example, a contemporaneous policy intervention that occurs at the same time as ECC may confound the results (in education, this can be a real threat due to multiple policy changes). This is not the case for RCT designs. Therefore, the impact results from the ITS analyses in the secondary analyses are treated more cautiously than the results from the more rigorous randomized controlled trials.

In a case control design (CCD), participants are identified with a specific intervention or outcome and compared with a matched control group of participants without the intervention or outcome. In this case the KS1 outcomes for schools already implementing ECC are compared with the KS1 outcomes for matched control schools. As with the ITS design, the CCD provides a mechanism for establishing a causal link between ECC and KS1 outcomes, but due to the limitations associated with matching, the causal link is not as strong as that provided by the more rigorous RCT design. This is because controls have not been randomly allocated and there is a real danger that selection bias will affect the results (because of the possibility that control schools will have subtle but important differences that may affect outcomes). Therefore, the results from the CCD analyses in the secondary analyses are treated more cautiously than the results from the more rigorous randomized controlled trials.

The secondary analyses use data from all of the intervention children in the Every Child Counts 2008-9, 2009-10 cohort schools (with the exception of the schools taking part in Trial 2), data from all of the children in these schools not exposed to the Every Child Counts intervention, historical data from the same schools and data from matched comparison

schools derived from the National Pupil Database (NPD). We assessed the impact of one-to-one delivery of the ECC intervention compared with control schools and children (not receiving the intervention) using 2009 KS1 outcomes.

1.3.3: Process evaluation

The process evaluation used a cross-sectional design. The key aims of the process evaluation were to understand how Numbers Count and the broader Every Child Counts initiative work and the implications of both for schools.

These aims were met by a number of inter-related research activities, including the following:

Classroom observations – these looked both at actual NC teaching as well as the wider impacts including re-integration (through whole class observations). A wide range of interviews were carried out with Numbers Count teachers (NCTs), existing Year 2 classroom teachers and teaching assistants (TAs).

Head teacher interviews – these were concerned with the school level impact of the programme, including recruiting and managing suitable NCTs, as well as long term resourcing issues and wider programme impacts.

Parental Interviews – these looked at the support the parents provided to their children and received from the schools.

Local Authority interviews – LAs have an important part to play in the delivery of the programme; this includes identifying suitable schools and having a key role in the appointment and support of teacher leaders (TLs).

Training / professional development – this looked at both the professional development for NCT and TLs. We carried out observations at the relevant events; individual and group interviews; and two on-line questionnaires.

Chapter 2: Impact: Trial 1

Key summary points

- The key findings are based on a rigorously designed and conducted pragmatic or real world randomized controlled trial (RCT) that used a waiting list design. The RCT evaluated the effectiveness of receiving Numbers Count (NC) compared with not receiving NC for attainment in mathematics.
- The primary analysis was based on data from 409 children.
- The children who received Numbers Count (intervention group) achieved statistically significantly higher average mathematics test scores than children who had not yet received NC (control group), using the PIM 6 mathematics test which was the primary outcome measure of short-term impact. The mean PIM 6 mathematics test score for the children receiving NC in the autumn term was 15.8 (SD 4.9) and for the control children who had yet to receive NC the score was 14.0 (SD 4.5). The effect size was 0.33 (95% CI 0.12 to 0.53), indicating strong evidence of a difference between the two groups (1.47 95% CI 0.71 to 2.23, $p<0.0005$) (Table 2-3). This is the equivalent to 7 additional weeks' improvement in the NC group, compared with the control group.
- We found no evidence that the results differed by any identifiable subgroup learner characteristics, including gender, free school meal status, age, and prior achievement.
- The Sandwell test was the secondary outcome measure of short-term impact. We were able to compare outcomes for intervention and control children on this test in January 2010 and again in April 2010. The effect sizes for this measure were 1.11 (95% CI 0.91 to 1.31) (January 2010) and 1.05 (95% CI 0.81 to 1.29) (April 2010). However, we were unable to minimise the threat to the reliability of this measure due to the potential for a number of biases: the test was specifically developed to be the diagnostic part of NC; it is a narrowly focused test; the testers knew whether the children they were testing had received NC or not.
- On the Performance Indicators in Primary Schools (PIPS) survey we found that Numbers Count improved attitudes to mathematics. However, it must be noted that the effect was marginally statistically significant and, given that several other measures on this survey were non-statistically significant, this finding might be the result of chance.

2.1: Introduction

Children who fall in the lowest 5% of the normal range in this age group (6-7 years) are predicted to struggle in numeracy attainment in future years. The programme Numbers Count (NC) was designed to improve these children's numeracy skills. In Trial 1 we

undertook a randomized controlled trial to address the key research question: What is the short-term impact of NC on numeracy skills compared with normal classroom practice?

2.2: Design and methods

We undertook a pragmatic ('real world') randomized controlled trial evaluating the effectiveness of NC in addition to normal classroom practice versus normal classroom practice alone for children's attainment in mathematics using a waiting list design. A pragmatic randomized trial is the best design to inform policy makers as it avoids selection bias, unlike non-randomized designs, and it gives more generalisable results than other randomized designs. It is also the most appropriate method for informing the conduct of the concurrent economic evaluation we undertook within the trial.

The trial was undertaken in authentic school and classroom settings in diverse geographical areas across England. The delivery of NC within the trial followed as closely as possible normal delivery of the programme. The primary outcome assessment was the PIM 6 mathematics test, which was undertaken blind to group allocation; the secondary outcome was the Sandwell test, which was not undertaken blind to group allocation. The PIM 6 test is a broader measure of mathematics attainment; the Sandwell test is more narrowly focused on the numeracy skills developed by NC.

Trial 1 assessed both the effectiveness of the Every Child Counts (ECC) intervention and the sustainability of the impact depending on the term of delivery. The participant schools for the trial were selected from the cohort of schools implementing the intervention for the *second* year in 2009-10. Schools had to have an accredited Numbers Count Teacher in order to be eligible to take part in the trial. The 12 eligible children within each school participating in the trial were individually randomly allocated to receive the intervention in one of three terms of delivery (autumn, spring or summer). We were able to assess the effectiveness of NC in January 2010 by using the post-test data from children due to receive the intervention in the second and third terms acting as controls for children who had already received NC in the first term.

The trial aimed to establish whether NC yielded superior results to normal classroom practice, and to measure the extent of the mean difference between children exposed to the intervention (and receiving normal classroom practice) compared with children not exposed to the intervention (but receiving normal classroom practice in mathematics and following the statutory content in the mathematics National Curriculum supported by the Primary National Strategy for numeracy).

The Protocol for Trial 1 (see 'Trial 1 Appendices' Appendix 1, in Torgerson et al, 2011c) emphasised the standardised training for delivery of the intervention and the standardised manual for implementation of the intervention which are normal practice, but also justified the ways in which implementation of the intervention was necessarily different in some minor details from standardised practice *for the purposes of the trial.*

Schools identified the children who were eligible to receive the intervention, and consent was obtained from the children and their parents to be involved in the trial, (specifically to undertake any additional testing that was necessary for the purposes of the trial including consent to take the wider outcomes tests). Once consent was checked and verified and the

baseline testing was completed, the schools contacted the Trial Co-ordinator either by telephone or by e-mail to access the randomization process which was undertaken by the York Trials Unit. This ensured unbiased allocation to trial arm.

2.2.1: Avoidance of bias

Randomization controlled for selection bias, temporal and regression to the mean effects, and the use of a secure, third party allocation system through the York Trials Unit ensured that random allocation could not be tampered with. In addition, because we used blinded assessment of outcome, this avoided the potential for ascertainment bias.

2.2.2: Outcome measures

Progress In Mathematics is a widely used commercial assessment from GL Assessment and has been developed (and re-standardised) from the NFER/Nelson 5-14 Mathematics assessment – which it has replaced. It is offered in a range of age related increments from 4 to 14 years. The PIM 6 version is appropriate for children six years of age. The assessment covers a wide range of mathematical skills and reflects the National Curriculum assessments at key stage 1 (KS1) and key stage 2 (KS2) (as well as the international assessment (TIMSS) for 9-10 year olds). The key areas assessed are: algebra; numbers and the number system; calculating; using and applying mathematics; shape, space and measures; handling data. PIM 6 can be administered to more than one child at once.

The Sandwell Early Numeracy Test was originally developed for use by the Sandwell Inclusion Support Service, and went on to be adopted by the Every Child a Chance Trust for use with Every Child Counts. Whilst the test is commercially available, its use outside ECC (and Sandwell) is, we understand, relatively limited. Two similar versions (A and B) are offered. The assessment covers national curriculum skills from P6 to level 2a, and concentrates on the following areas of numeracy: identification of numbers; oral counting; object counting; value and computation; language. These areas focus on number and largely coincide with the underlying approach of Numbers Count. Numbers Count focuses on number in the belief that this will lead to equivalent gains in other areas of mathematics (e.g., space and shape) (Numbers Count Handbook 2008 p11). As part of the NC programme the test is administered by the NC teacher, or other teachers/trained members of staff, on a one-to-one basis prior to the child starting the programme. The test is also administered on exit, and three and six months after the end of the programme by a link teacher.

ECC evaluation

Following discussions with the Steering Committee it was agreed that PIM 6 would be the outcome measure (assessment) for Trial 1 and Trial 2. This was for the following reasons: PIM 6 is a well recognised and reliable standardised test; it is not part of the Numbers Count programme and could therefore be administered independently of the programme; the evaluators could ensure that the people administering and marking the test were blinded to the groups (NC or control); it is a programme independent measure; it is a broad measure of mathematics achievement; it can be administered to more than one child at once (i.e., it is cost effective).

We have also reported results from the Sandwell assessments, as secondary outcome measures. It should though be noted that this testing was not undertaken independently; and the people administering and marking the tests were not blinded to the group allocations. Moreover, the test itself is a programme inherent measure, and assesses a narrower range of mathematic skills.

We have discussed below some of the key differences between the PIM 6 and Sandwell assessments and how they were administered.

Administration of tests

The Sandwell test was administered on a one-to-one basis by someone the child knew and it was not time limited – i.e., the child was allowed to take as long as they wanted to complete the test. The PIM 6 tests were administered in a controlled way by people unknown to the children (mostly OFSTED inspectors). This was with groups of children and the testing followed a strict protocol. There may also have been a practise effect applicable to the Sandwell test because it formed the pre-test, the post-test and the follow-up tests.

Programme inherent or independent measures

The Sandwell test can be considered a programme inherent measure, in that it is used for the initial assessment and as a diagnostic tool in the programme. It is also used at the end of the programme (and beyond). It follows that much of the teaching will be determined by weaknesses identified by the Sandwell test and therefore we would expect particularly good progress to be made in these areas. PIM 6 was used independently of the diagnostic or teaching element of the programme. Slavin and Madden (2008), in their comparison of studies included in the What Works Clearinghouse, found that the average effect size for studies of mathematics interventions using treatment (programme) inherent measures (as in the case of Sandwell) was +0.45, whereas for treatment independent measures (as in the case of PIM 6) ES was -0.03.

Narrow or broad measures

Related to the above, the Sandwell test provides a narrower assessment of mathematical skills compared to PIM 6 (on the assumption that the skills not taught (e.g., space and shape) will also improve). This is relevant because the NC programme works on the principle that equivalent gains in other areas of mathematics will be made.

Conclusions

The Sandwell test has been used to aid the diagnostic process in the programme, but it does not provide a good measure of programme impact for the reasons outlined above. The impact finding of an effect size of 0.36 on PIM 6 (moderately large, statistically significant) should be recognised as a reliable and robust benchmark of short term programme impact for this evaluation.

Pre-testing and post-testing

All 12 ECC children in each school had a pre-test (Sandwell A), *after which* they were randomly allocated by the York Trials Unit into three groups: Group 1 received NC in the autumn term (term 1), Group 2 received NC in the spring term (term 2); and Group 3 received NC in the summer term (term 3). All children were post-tested using the primary outcome PIM 6 (independent test) at the beginning of the spring term (January testing). All children were all post-tested using the secondary outcome, Sandwell test, at the end of the first term (Sandwell B), at the end of the second term (Sandwell A) and there was a final post-test at the end of the third term (Sandwell B). We used the results from the KS1 (literacy and numeracy) tests in May as a third outcome, although there are a number of limitations in the use of this measure. The PIM 6 (administered in January 2010) was the primary outcome measure for the main randomized comparison between intervention and control children. This test was undertaken and marked independently and blind to group allocation. The secondary outcome was the Sandwell test, which was not undertaken or marked blind to group allocation.

Table 2-1: Trial 1 Testing regime

Pupils	Baseline September 2009	January testing December 2009/January 2010	April testing March 2010/April 2010	July testing July 2010
4 children receiving NC in autumn term 2009	Sandwell A (Entry)	Sandwell B (Exit) INDEPENDENT TEST Progress in mathematics 6 Wider outcomes assessments (PIPS, SDQ) (optional)	Sandwell A (3 month)	Sandwell B (6 month)
4 children receiving NC in spring term 2010	Sandwell A	Sandwell B (Entry) INDEPENDENT TEST Progress in Mathematics 6 Wider outcomes assessment (PIPS, SDQ) (optional)	Sandwell A (Exit)	Sandwell B (3 month)
4 children receiving NC in summer term 2010	Sandwell A	Sandwell B INDEPENDENT TEST Progress in Mathematics 6 Wider Outcomes assessment (PIPS, SDQ) (optional)	Sandwell A (Entry)	Sandwell B (Exit)
Normal practice and required by evaluation				
Additional testing/assessment required by evaluation				

Note: All 12 pupils were allocated at random to autumn, spring or summer term. Schools could choose to withdraw any pupils who could be potentially 'harmed' by randomization before randomization was conducted; these pupils did not receive ECC unless the schools created additional slots. All normal practice entry tests were conducted by the Numbers Count Teachers or a trained Teaching Assistant. All normal practice exit and follow up tests were conducted by the Link teacher, the NC teacher or a trained TA.

2.2.3: Economic evaluation

We believe it is very important to include a trial based economic evaluation. We undertook an economic evaluation within Trial 1. We did this by collecting data on the incremental cost of NC. We compared these costs with the additional gains from the programme to assess the cost per extra child who was working at the equivalent of Level 2c in mathematics at KS1 as estimated by the outcome measurement in Trial 1 in January (i.e., 5 or 6 months before normal assessment at KS1). (Please see 'Trial 1 Appendices' Appendix 5 for the protocol, Torgerson et al, 2011c; please see Chapter 6 for the economic evaluation.)

2.2.4: Wider impact

In addition to the assessment of impact on numeracy abilities, we measured the following variables in order to assess the wider impact of the intervention in exploratory analyses:

(a) Attention/behaviour/mental health (Strengths and Difficulties Questionnaire (SDQ) Goodman (2001) teacher/parent scale) (sample of children, assessment not blinded);

(b) Attitudes to mathematics, literacy and school (Performance Indicators in Primary Schools, PIPS, survey (all children, assessment blinded).

Table 2-2: Trial 1 Outcome measures

Primary outcome	**Secondary outcomes**	**Exploratory outcomes**
PIM 6 January (January testing)	Sandwell B December (January testing) Sandwell A March (April testing) Sandwell B July (July testing)	PIPS Quiz January (January testing) SDQ December (January testing)

Table 2.2 gives the primary, secondary and exploratory outcomes with timelines.

A paper based survey was developed which sought factual information about teachers' experience and qualifications and included a log for the teachers to record each child's participation in NC. This information was used: a) as a check for fidelity of implementation (but note this was by self-report, not independent observation, and conclusions derived from results take this limitation into account); and b) to gather information on variables for subgroup analyses.

Table 2-3: Trial 1 Data collection table

	PIM 6	Sandwell A	Sandwell B	PIPS Quiz	SDQ	Pupil Log	Teacher Survey
September 2009 (Baseline)		X					X
December 2009 (January testing)			X		X	X	
January 2010 (January testing)	X			X			
March 2010 (April testing)		X				X	
April 2010 (April testing)							
July 2010 (July testing)			X			X	X

Table 2.3 gives the data collection information for the trial, with timelines.

2.2.5: Sample size calculation

The power calculations are based upon the following data. We expected the intervention group to improve by 1.25 standard deviations or greater compared with the pre-test value and we wished to detect a marginal increase of 0.25 compared with the waiting list controls. We also assumed a pre-test post-test correlation of at least 0.70. To have at least a 95% chance of observing such a difference we needed approximately 600 children in our sample given a randomization ratio of one-to-two (i.e., at the end of the first term 8 children were in the control group and 4 were in the intervention group). To recruit this number required a total of 50 schools. We anticipated that we would recruit 50 schools which would give 95% power to observe 0.25 of an effect size.

2.2.6: Statistical analysis

All analyses were conducted on an intention to treat basis. Consequently any children who crossed over from either study arm were analysed as per their randomized allocation. Analyses were conducted in Stata using 2-sided significance tests at the 5% significance level. All baseline data were summarised by treatment group and presented descriptively. No formal statistical comparisons were undertaken for the baseline data. The primary outcome was the PIM 6 mathematics test. The scores on the PIM 6 were summarised descriptively (mean and standard deviation) by allocated group. Linear regression was used to compare the two groups with adjustments made for the potential clustering within schools using the Huber-White sandwich estimator (robust standard errors). The outcome modelled was the PIM 6 score and the model included age, gender, free school meal status, Sandwell A test score (pre-test) and group allocation. This analysis was repeated for the secondary outcome which was the Sandwell test.

The main analysis compared the children receiving the NC intervention in the autumn term with the children who were allocated to receive the NC intervention in the spring or summer terms.

The anonymity of all schools, children and teachers was preserved for all analyses and there is no presentation or comparison of the results from individual schools or teachers. Subgroup analyses assessed the effectiveness of the intervention for children with different learner characteristics (gender, free school meal status etc) and depending on a number of teacher characteristics. Learner and teacher characteristics were obtained through pupil logs and teacher surveys. Our design dealt with the possibility that the experimental group may have looked artificially good immediately after the intensive one-to-one teaching in part by comparing the first cohort to the third cohort on the April assessment (although this was only possible using the secondary outcome measure). In addition, by comparing the first to the second cohort at that time we also checked to see whether, in fact, there was a one-time bump in scores immediately after intensive tutoring.

2.2.7: Wider impact

For the wider impact assessments we compared the mean score for the intervention group with the mean score for the control group using linear regression with adjustments made for the potential clustering within schools using the Huber-White sandwich estimator (robust standard errors). The individual outcomes modelled were the PIPS total score, PIPS attitude to mathematics, PIPS attitude to reading, PIPS attitude to school, SDQ parent score and SDQ teacher score. For each outcome the model also included age, gender, whether the child was receiving free school meals and group allocation. The appropriateness of the method of linear regression and the confidence intervals and tests of significance depends on the assumption that the residuals are normally distributed. There were obvious departures from the assumption for the PIPS outcomes; thus to test the robustness of the results a sensitivity analysis was also undertaken using ordinal regression.

There were a number of missing responses for the three domains (attitude to mathematics, reading and school) of the PIPS measure. Additional analyses were undertaken to explore the impact of missing data. For each domain if one of the two responses was missing, then the missing response was replaced with the other response. If both of the responses were missing for the domain, then the response was excluded from the analysis.

2.2.8: Key stage 1 outcomes

Logistic regression was used to compare the two groups with adjustments made for the potential clustering within schools using the Huber-White sandwich estimator (robust standard errors). The individual outcomes modelled were the key stage 1 (KS1) mathematics, reading, writing and science scores and the models included age, gender, whether the child was receiving free school meals and group allocation. For the KS1 mathematics analysis, Sandwell A test score (pre-test) was also included in the model.

2.2.9: Quality assurance procedures for designing and reporting RCTs: the CONSORT guidelines

We designed, conducted and reported the trial using the CONSORT guidelines or statement (Altman et al, 2001). CONSORT was developed by a collaboration of medical journal editors and leading trial methodologists to ensure that medical trials were conducted and reported to the highest standards. CONSORT has recently been adopted by leading psychological journals and some educational journals.

Applying the CONSORT guidelines to the design of trials ensures that key methodological criteria, such as the method of randomization, are explicitly reported. This allows the reader to judge whether or not the trial is of high quality. Because we designed Trial 1 around the CONSORT statement this ensures that it was conducted and reported to the highest quality standards.

2.2.10: Research ethics and data management

We submitted our research plan (Protocol) for the trial to the University of York Humanities and Social Science Ethics Committee for ethical approval. Data processing and management abided by current data protection regulations. All data were stored on secure servers that are password protected. All electronic data can be held indefinitely. We used the SRA research ethics framework (see 'Trial 1 Appendices' Appendix 3 for full data protection issues, Torgerson et al, 2011c). We received approval for our protocol from DfE and the Steering Group. The trial protocol includes dates approval was received from the University of York HSSEC, the DfE and the Steering Group. All trial, ethics and testing protocols, information and consent forms, and all trial school correspondence templates are included in the Appendices to this document (Torgerson et al, 2011c).

2.3: Results

2.3.1: School progress through the trial

Data from 44 schools are included in the analysis for this report.

The participation and progress of schools and children in Trial 1 are shown in Figure 2.1. Of the 68 schools approached to take part and invited to a recruitment conference, 53 schools agreed to take part in the trial. After giving consent, 8 schools had to withdraw from the trial as their trained Numbers Count Teacher (NCT) was replaced with a new NCT (and only schools with accredited NCTs were eligible to take part in the trial). 44 of the remaining 45 schools completed pre-tests (Sandwell A September 2009 testing) for almost all children selected for the trial in their school. Individual randomization of all children to autumn, spring or summer delivery of the intervention was undertaken by the end of September 2009. Four children were tested late (after randomization) because they were absent at the beginning of the autumn term. Two children were never pre-tested (Sandwell A September 2009 testing). The decision was taken to allow the one remaining school to drop out of the trial in October 2009 as the NCT had not been able to conduct the pre-test with any of the 12 selected children. Consequently, all children in 44 schools were randomized and remained in the trial for the duration of the autumn term. One school decided to withdraw from the trial at the end of the autumn term because of the increase in paperwork due to being involved in the trial. A further school also withdrew at the end of the autumn term as the NCT retired and the school did not intend to appoint a replacement. Thus, 42 schools remained in the trial as of January 2010. One school had to withdraw from the trial at the end of the spring term as their NCT left and was not replaced for the summer term. Consequently, 41 schools remained in the trial at the end of the study.

2.3.2: Pupil progress through the trial

Each school was asked to identify 12 children to take part in the trial. 41 of the 44 randomized schools selected 12 children. One school selected 11 children and a second school selected 10 children because the schools could not obtain parental consent in time before the randomization needed to be undertaken to enable teaching with the other children to begin (children without parental consent were taught by the teacher in the gaps, once parental consent for NC only had been gained, but these children were not in the trial). One school selected 9 children because they only had 9 children eligible for the intervention. In total 522 children with parental consent took part in the trial. Of these, 18 children from 8 schools were randomized to the spring or summer term only. They were excluded from being randomized to the autumn term delivery of NC, either because they were also receiving Reading Recovery (RR) (16), or because they moved between schools at the beginning of term (2). These children were excluded from all analyses involving children allocated to the autumn term. One child who was randomly allocated to receive NC in the autumn term actually received NC in spring term. By the end of the autumn term 13 children had left the trial because they had moved to another school (4 from Group 1, 4 from Group 2, 5 from Group 3). At the end of the autumn term 509 children remained in the trial.

At the beginning of the spring term 24 children were withdrawn because their schools withdrew. Two further children left the trial at the beginning of the spring term because they moved school (1 from Group 2, 1 from Group 3). One child who was randomly allocated to receive NC in the spring term was in hospital during this time, and the school taught a child who was randomly allocated to summer term delivery of NC instead. During the spring term an additional 8 children left the trial because they moved to another school (1 from Group 1, 2 from Group 2 and 5 from Group 3). Thus, at the end of the spring term 475 children remained in the trial.

At the beginning of the summer term 12 children were withdrawn because their school withdrew. Three further children left the trial at the beginning of the summer term because they moved school (1 from Group 1, 1 from Group 2, 1 from Group 3). The child randomly allocated to receive NC in the spring term who had been in hospital during that term received NC in the summer term. The child randomly allocated to the summer term who had received NC in the spring term instead received normal classroom practice in the summer term. Two further children who were randomly allocated to receive NC in the summer term did not receive NC as their school felt each child was performing well and would no longer benefit from the intervention. Data for one of these children were still collected and analysed (in the ITT analysis). One further child randomly allocated to NC in the summer term was absent at the beginning of the term and the school taught another child in their place. However data for this child were still collected and analysed (in the ITT analysis). During the summer term 6 children left the trial because they moved school (1 from Group 1, 4 from Group 2, 1 from Group 3). Thus, at the end of the summer term 454 children remained in the trial.

Figure 2-1: Trial 1 CONSORT diagram

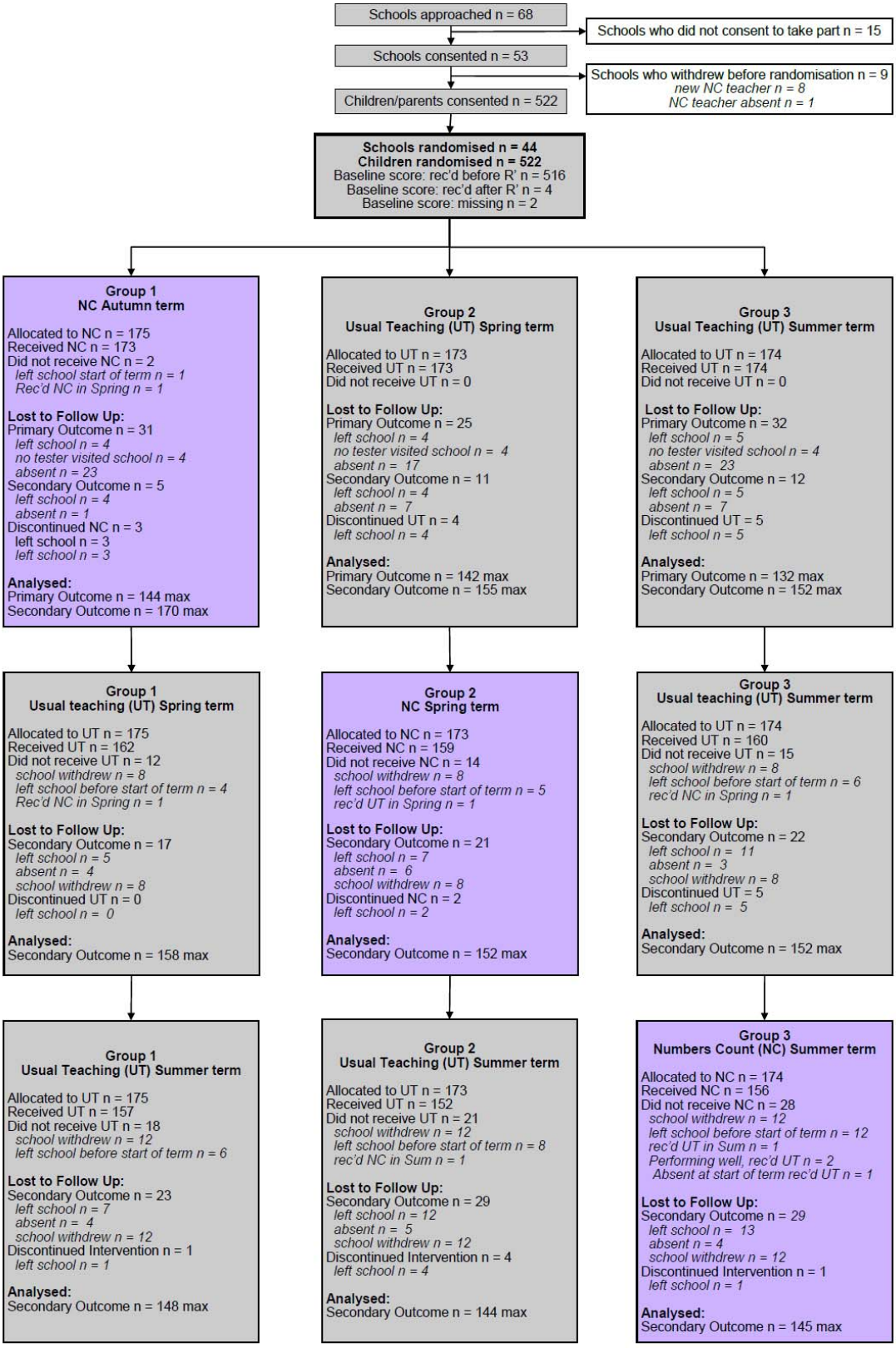


Table 2-4: Trial 1 Number of data points returned

	PIM 6	Sandwell A	Sandwell B	PIPS Quiz	SDQ Parent	SDQ Teacher	Pupil Log	Teacher Survey
September 2009 (Baseline)		520						38
December 2009 (January testing)			494		112	277	156	
January 2010 (January testing)	434			434				
March 2010 (April testing)		462					158	
April 2010 (April testing)								
July 2010 (July testing)			441				139	36

Maximum=522; for pupil log autumn max=175, spring max=173, summer max=174; for teacher survey max=38.

Protocol deviations

- One child randomly allocated to receive NC in the autumn term received it in the spring term.
- One child randomly allocated to receive NC in the spring term was in hospital during this time; the school replaced this child with another child randomly allocated to summer term.
- Two children randomly allocated to receive NC in the summer term did not do so as their school felt each child was performing well and would no longer benefit from the intervention.
- One child randomly allocated to NC in the summer term was absent at the beginning of the term and the school taught another child in their place.
- Four children were tested late (after randomization) and 2 children were never pre-tested because they were absent at the beginning of the autumn term.
- Eight children from 1 school were pre-tested using the old version of the Sandwell test.
- Nine children from 1 school were tested using the Sandwell A test rather than the Sandwell B test in the January testing point.
- Two children from 2 schools were tested late (secondary outcome) in the January testing period.
- Four children from 4 schools were tested late (secondary outcome) in the April testing period.

Due to the small number of protocol deviations and the fact that we adhered to an intention-to-treat data analytic plan, we do not think that the deviations introduced a source of bias into the analysis.

2.3.3: Baseline characteristics

Table 2.5 gives the baseline characteristics of children included in Trial 1. The characteristics are summarised by the term of delivery. 18 children were randomized to receive NC in the spring or summer terms only and have been excluded from the summaries below. As expected, randomization resulted in all groups having similar characteristics.

Table 2-5: Trial 1 Baseline characteristics

Characteristics	Autumn	Spring	Summer
Age, mean (SD)	6.4 (0.3) [n=173]	6.5 (0.3) [n=165]	6.4 (0.3) [n=162]
Sandwell A, mean (SD)	28.2 (8.4) [n=174]	26.7 (8.3) [n=165]	27.0 (8.7) [n=163]
Free school meal, n (%)	86 (50.9) [n=169]	63 (39.1) [n=161]	76 (48.7) [n=156]
Gender (females), n (%)	69 (39.7) [n=174]	65 (39.4) [n=165]	72 (44.2) [n=163]

Between randomization and post-test of the primary outcome measure approximately 86 (17%) of children were 'lost to follow-up' or withdrawn. There were a number of reasons for those lost to follow-up, including absence from school during the post-test (this was especially a problem as many children were due to be tested during the bad, snowy weather in January 2010). However, we do not believe that this attrition is likely to have introduced bias as the proportion missing from each group was similar (i.e., 31 (18%), 24 (14%) and 31 (19%) for autumn, spring and summer respectively) and there did not appear to be any systematic reasons for the attrition that would have been related to group allocation.

2.3.4: Primary outcome

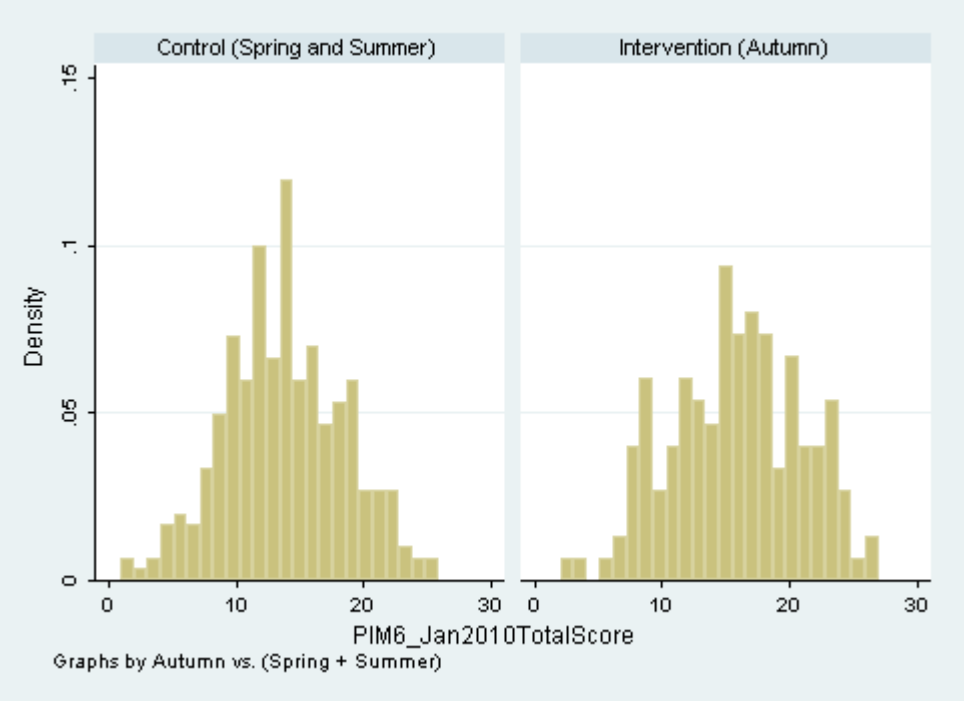
The primary outcome measure was the PIM 6, which was undertaken and marked blind to group allocation by independent testers to minimise the threat to reliability of the test due to outcome ascertainment bias.

Table 2-6: Trial 1 Summary of primary outcome measure

Outcome	Intervention	Control	Estimate(95% CI)*	Effect size
PIM 6, mean (SD)	15.8 (4.9) [n=144]	14.0 (4.5) [n=274]	1.47 (0.71 to 2.23) [n=409]	0.33 (0.12 to 0.53)

* Analyses were adjusted for baseline Sandwell A test scores, age, gender, free school meals and the clustering within schools. Analyses exclude children who could not be randomized to autumn term.

Figure 2-2: Histogram showing the PIM 6 scores for the intervention and control groups



The mean PIM 6 score for the children receiving NC in the autumn term was 15.8 (SD 4.9) and for control children who had yet to receive NC was 14.0 (SD 4.5). The effect size was 0.33 (95% CI 0.12 to 0.53), indicating strong evidence of a difference between the two groups (1.47 95% CI 0.71 to 2.23, p<0.0005) (Table 2.6). This result indicates that children who received NC scored significantly higher on the PIM 6 mathematics test compared with children who had not received NC.

2.3.5: Secondary outcome

The secondary outcome measure was the Sandwell test (A or B depending on time of assessment) (see Table 2.7), which was undertaken and marked by NCTs or class teachers or teaching assistants, who were not blind to group allocation.

Table 2-7: Trial 1 Descriptive statistics of secondary outcome measures over time

Assessment	Autumn		Spring		Summer	
	N	Mean (SD)	N	Mean (SD)	N	Mean (SD)
Sandwell A (Sep)	174	28.2 (8.4)	165	26.7 (8.3)	163	27.0 (8.7)
Sandwell B (Jan)	170	45.0 (11.1)	155	32.3 (9.9)	152	32.7 (10.6)
Sandwell A (Apr)	158	48.7 (10.6)	147	48.2 (12.1)	144	37.0 (11.0)
Sandwell B (Jul)	152	52.8 (11.4)	139	51.9 (13.1)	137	50.9 (12.5)

*Excludes children unable to be randomized to the Autumn term

Figure 2.3 shows the trajectory of children in terms of their attainment in the Sandwell tests by the time of intervention. The figure shows that children improved their Sandwell test scores once they received Numbers Count.

Figure 2-3: Trial 1 Summary of secondary outcome measure over time

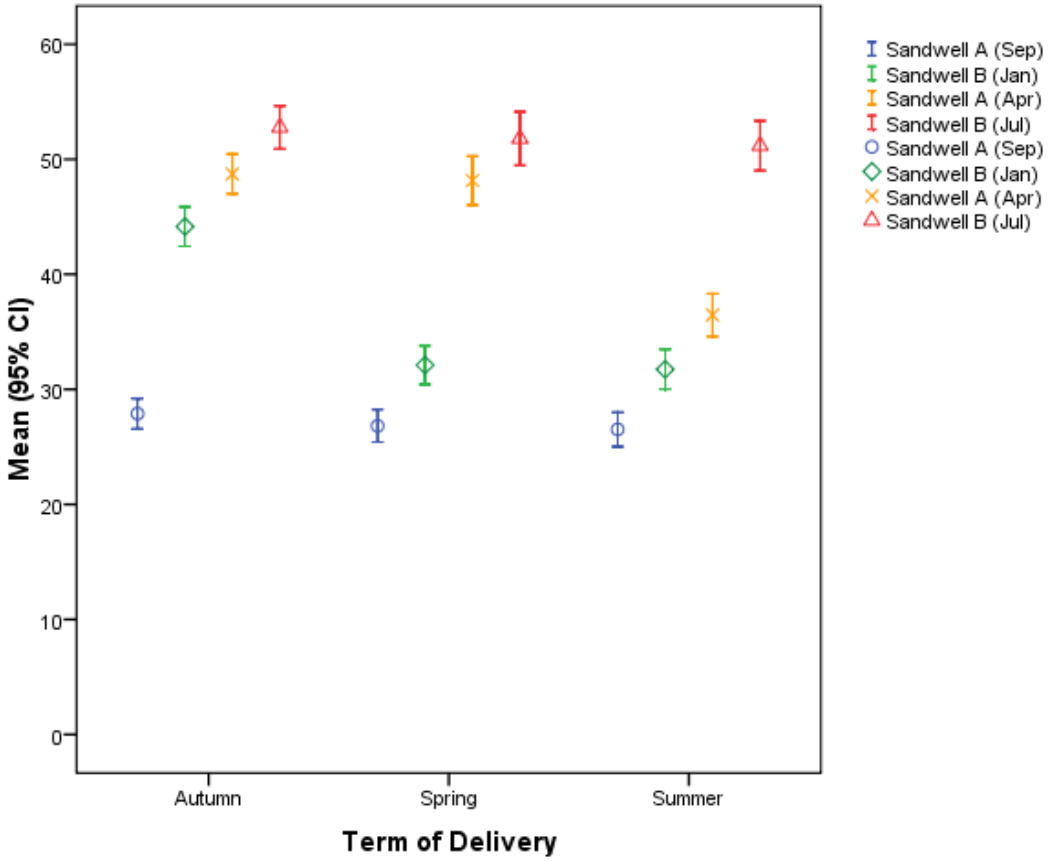


Table 2-8: Trial 1 Secondary outcome, Intervention vs. waiting list control

Outcome	**Autumn**	**Spring + Summer**	**Estimate (95% CI)** *	**Effect size**
Sandwell B (Jan), mean (SD)**	45.0 (11.1) [n=170]	32.5 (10.2) [n=307]	11.3 (9.6 to 13.1) [n=464]	1.11 (0.91 to 1.31)
Outcome	**Spring**	**Summer**	**Estimate (95% CI)** *	**Effect size**
Sandwell A (Apr), mean (SD)	48.0 (12.2) [n=152]	36.8 (10.9) [n=152]	11.4 (9.5 to 13.3) [n=295]	1.05 (0.81 to 1.29)

*Analyses were adjusted for baseline Sandwell A test scores, age, gender, free school meals and the clustering within schools

** Analysis excludes children who could not be randomized to autumn term.

The Sandwell results were very similar for the January and April assessments. The effect sizes for this measure were 1.11 (95% CI 0.91 to 1.31) and 1.05 (95% CI 0.81 to 1.29). However, we were unable to minimise the threat to internal reliability of this measure due to the potential for outcome ascertainment bias. Therefore, the higher effect size for this

measure may be due to bias rather than any differences in sensitivity of the measurement tool compared with the PIM 6.

2.3.6: Comparison of outcomes by term of delivery

We compared children who received NC in the autumn term to the children who received NC in the spring term on the April assessment using the secondary outcome measure (Sandwell A). We repeated this analysis after the July testing using the secondary outcome measure (Sandwell B), this time comparing outcomes for all three terms. The results of these analyses are presented in Table 2.9.

Table 2-9: Trial 1 Secondary outcome measures for delivery of intervention

Outcome	**Autumn**	**Spring**	**Estimate (95% CI)** *	**Effect size**
Sandwell A (Apr), mean (SD)	48.7 (10.6) [n=158]	48.2 (12.1) [n=147]	0.51 (-1.17 to 2.18)	0.05 (-0.18 to 0.27)
Outcome	**Autumn**	**Spring**	**Summer**	**P-value**
Sandwell B (Jul), mean (SD)	52.8 (11.4) [n=152]	51.9 (13.1) [n=139]	50.9 (12.5) [n=137]	0.78

*Analyses were adjusted for baseline Sandwell A test scores, age, gender, free school meals and the clustering within schools [Sandwell A N=296; Sandwell B N=415]

The mean Sandwell A (April) score for the children receiving NC in the autumn term was 48.7 (SD 10.6); the score for children receiving NC in the spring term was 48.2 (SD 12.1). The effect size was 0.05 (95% CI -0.18 to 0.27), indicating little or no evidence of a difference between the two groups (0.51 95% CI -1.17 to 2.18, p=0.54) (Table 2.9). The mean Sandwell B (July) score for the children receiving NC in the autumn term was 52.8 (SD 11.4); the score for children receiving NC in the spring term was 51.9 (SD 13.1); and the score for children receiving NC in the summer term was 50.9 (SD 12.5). The results indicate little or no evidence of a difference between the three groups (p=0.78) (Table 2.9). The results of both analyses indicate there was little or no evidence of a one-time bump in scores immediately after intensive tutoring. This means that there was an increase in scores immediately following the intervention and this does not appear to decrease over the remainder of the follow up period.

2.3.7: Exploratory analysis

For the primary outcome measure (PIM 6) we also explored whether children responded differently to NC based upon a number of pupil and teacher characteristics. From the analysis, we found no evidence that children responded differently to NC based upon any of the pre-specified interactions: pre-test scores (p=0.18), age (p=0.41), gender (p=0.97), free school meals (p=0.83), number of years' teaching experience (p=0.42), or NCTs' highest educational achievement (p=0.61). (Note: full results from the pupil logs and teacher surveys are presented in Appendices to the report, see Appendix 34 and Appendix 35, Trial 1 Appendices, Torgerson et al, 2011c).

2.3.8: Wider impact analysis

The results indicate no evidence of a difference between the two groups for the PIPS total score (0.21 95% CI -0.36 to 0.77, p=0.46), attitude to reading (0.10 95% CI -0.19 to 0.39, p=0.50), attitude to school (-0.06 95% CI -0.30 to 0.18, p=0.60), SDQ teacher score (0.08 95% CI -2.57 to 2.73, p=0.95) and SDQ parent score (-0.67 95% CI -3.79 to 2.45, p=0.65; Table 2.10). However, the findings from the PIPS attitude to mathematics questions indicate evidence of a difference between the two groups (0.21 95% CI 0.02 to 0.40, p=0.03). This result indicates that children who received NC scored significantly higher on the questions relating to their attitude to mathematics compared with children who had not received NC. However, it must be noted that the effect was marginally statistically significant and, given that several other measures were non-statistically significant, this finding may be the result of chance.

Table 2-10: Trial 1 Wider impact assessment

Assessment	Autumn		Spring + Summer		Estimate (95% CI)
	N	Mean (SD)	N	Mean (SD)	
PIPS					
Total	132 *143*	15.3 (2.9) *15.2 (2.9)*	249 *270*	15.0 (2.6) *14.9 (2.6)*	0.21 (-0.36 to 0.77) *0.22 (-0.35 to 0.79)*
Mathematics	139 *143*	5.1 (1.1) *5.1 (1.1)*	259 *272*	4.8 (1.1) *4.8 (1.2)*	0.21 (0.02 to 0.40) *0.21 (0.02 to 0.40)*
Reading	138 *143*	5.1 (1.3) *5.0 (1.3)*	265 *271*	4.9 (1.3) *4.9 (1.3)*	0.10 (-0.19 to 0.39) *0.05 (-0.24 to 0.35)*
School	140 *143*	5.1 (1.2) *5.1 (1.3)*	268 *272*	5.2 (1.1) *5.1 (1.1)*	-0.06 (-0.30 to 0.18) *-0.05 (-0.30 to 0.20)*
SDQ					
Teacher	39	12.6 (6.2)	67	13.4 (7.8)	0.08 (-2.57 to 2.73)
Parent	71	11.5 (7.4)	143	11.4 (6.4)	-0.67 (-3.79 to 2.45)

Values in italics are summaries when adjustments are made for missing data

Sensitivity analysis

The analysis for the PIPS assessment was repeated using ordinal regression to explore the robustness of the findings. The conclusions drawn from the sensitivity analysis were consistent with the findings from the wider impact analysis.

2.3.9: Key stage 1 outcomes

Table 2.11 summarises the KS1 levels for each subject by the term of delivery of NC.

Table 2-11: Trial 1 Overall Key stage 1 outcomes

	Randomized term		
	Autumn N (%)	**Spring N (%)**	**Summer N (%)**
Mathematics			
W	3 (2%)	5 (3.6%)	2 (1.5%)
1	39 (26.4%)	42 (30.2%)	53 (38.7%)
2C	58 (39.2%)	45 (32.4%)	50 (36.5%)
2B	30 (20.3%)	34 (24.5%)	24 (17.5%)
2A	16 (10.8%)	11 (7.9%)	6 (4.4%)
3	2 (1.4%)	2 (1.4%)	2 (1.5%)
Reading			
W	7 (4.7%)	8 (5.8%)	11 (8%)
1	63 (42.6%)	66 (47.5%)	61 (44.5%)
2C	44 (29.7%)	24 (17.3%)	33 (24.1%)
2B	24 (16.2%)	32 (23%)	27 (19.7%)
2A	7 (4.7%)	5 (3.6%)	4 (2.9%)
3	3 (2%)	4 (2.9%)	1 (0.7%)
Writing			
W	12 (8.1%)	14 (10.1%)	18 (13.1%)
1	74 (50%)	68 (48.9%)	63 (46%)
2C	40 (27%)	34 (24.5%)	34 (24.8%)
2B	16 (10.8%)	19 (13.7%)	20 (14.6%)
2A	5 (3.4%)	4 (2.9%)	2 (1.5%)
3	1 (0.7%)	0 (0%)	0 (0%)
Science			
W	2 (1.4%)	2 (1.5%)	3 (2.3%)
1	47 (33.6%)	53 (40.2%)	52 (40.3%)
2	83 (59.3%)	73 (55.3%)	72 (55.8%)
3	8 (5.7%)	4 (3%)	2 (1.6%)

Table 2.12 displays the number of children scoring level 2c or above for KS1 mathematics, reading, writing, and science outcomes.

Table 2-12: Trial 1 Summary of the key stage 1 outcomes

Outcome	Autumn	Spring	Summer	P-value*
Mathematics	106/148 (71.6)	92/139 (66.2)	82/137 (59.9)	0.34
Reading	78/148 (52.7)	65/139 (46.8)	65/137 (47.4)	0.66
Writing	62/148 (41.9)	57/139 (41.0)	56137 (40.9)	0.98
Science	91/140 (65.0)	77/132 (58.3)	74/129 (57.4)	0.30

* Analyses were adjusted for age, gender, free school meals and the clustering within schools

The results presented in table 2.12 highlight the lack of evidence of an overall difference among the three terms for any of the KS1 outcomes assessed: mathematics (p=0.34), reading (p=0.66), writing (p=0.98), and science (p=0.30).

2.4: Discussion

We report the results of a randomized controlled trial (Trial 1) comparing one-to-one Numbers Count teaching with normal classroom practice. Our results demonstrate a statistically significant effect of Numbers Count on our primary, independently marked, mathematics test. The effect size of 0.33 is reasonable for a pragmatic field trial (i.e., where the intervention was given to children in a 'normal' school setting).

We also found that, in line with the process evaluation findings, NC had a positive impact on the attitudes of the children towards mathematics (although the effect was marginally statistically significant). However, we could not find any evidence of improved attitudes to school more generally, and this is in contrast to the feedback we received as part of the process evaluation.

We can look at the results in other ways which can help with interpretation of the educational significance of the effect size difference we noted. If we assume a bench mark of the mean score of the control group then the results are consistent with between 12-16% more children in the intervention group getting a score higher than the mean score of the control group.

For our secondary outcome, the Sandwell test, the effect size is larger in both the January testing comparison and the April testing comparison. However, this test was administered by teachers who were aware of the group membership of the children. Therefore, it is possible that the results of this test may be biased due to outcome ascertainment (post-test measure) not being blind to group allocation. It is also possible that some of the difference in effect size between the primary and secondary outcomes could reflect different mathematical concepts being measured in the primary outcome test and secondary outcome test, with the Sandwell test focusing on assessment of use of number and the PIM 6 test assessing mathematical abilities more broadly.

The trial has a number of key strengths. The randomization used a specifically written software programme from the York Trials Unit which maintained its security. Observer bias was eliminated in the primary outcome measure through the use of independent testers who were unaware of the group allocation of the children being tested. The completed tests were marked by independent testers unaware of the group allocation. Finally, the researchers from the Universities of York, Durham and Birmingham were independent.

The trial does, however, have a number of limitations. We had a significant attrition rate for our primary outcome measure, which was in part, we believe, due to the inclement weather during testing. However, because attrition was virtually identical between the intervention and control groups and because there did not appear to be any systematic reason for the attrition we do not think significant bias was introduced. Our actual sample size was somewhat lower than we had anticipated (i.e., 418 rather than 600); however, the effect size we observed (0.33) was somewhat greater than anticipated in the original sample size calculation. Consequently, there was little loss of power in our study. Because of the short-term nature of the trial we could not look at the longer term effects of NC. Ideally a randomized trial with a longer-term follow-up would be necessary to see whether or not the intervention could have 'washed out' over time with the control children 'catching up' using 'normal' classroom teaching. We could not disentangle the effect of one-to-one teaching *per se* from NC one-to-one teaching. It may be that offering a different one-to-one mathematics intervention could have had similar effects. Comparing NC with other mathematics programmes would be fruitful, as would be a study with long term follow-up, to ensure the effects of the intervention are maintained over time.

2.5: Conclusions

In Trial 1 we can conclude that one-to-one NC is a more effective method of improving numeracy skills among children with the lowest performance in mathematics, compared with normal classroom practice, when assessed after a term of the intervention. The effect size is in line with pragmatic field trials of effective interventions in education. We cannot say whether this effect persists beyond one term.

Chapter 3: Impact: Trial 2: Pairs

Key summary points

- The key findings are based on a robustly designed and conducted randomized controlled trial evaluating the effectiveness of Numbers Count in comparison to adapted Numbers Count delivered to pairs of children for attainment in mathematics.

- The primary analyses are based on data from 66 children (January 2010 testing) and 63 children (April 2010 testing).

- The independent test that was used to measure attainment in mathematics was the PIM 6 mathematics test. Testing was undertaken in January 2010 and April 2010. This was the primary outcome measure of the short-term impact of NC compared with adapted NC delivered to pairs of children. The secondary outcome was the Sandwell mathematics test, which was not independently administered.

- In January 2010, on the primary outcome measure (PIM 6), there were no statistically significant differences in the PIM 6 mathematics test scores between children receiving Numbers Count and children receiving adapted Numbers Count in groups of two, suggesting that the interventions have similar levels of effectiveness. The mean score for the children receiving Numbers Count was 15.5 (SD 3.5) and for the children receiving adapted NC in pairs was 17.1 (SD 5.0). The effect size was 0.30 (95% CI -0.31 to 0.91). The results demonstrate that children who received adapted NC in pairs scored slightly higher, although not statistically significantly higher, on the PIM 6 mathematics test compared with children who had received NC individually (1.4 95% CI -0.6 to 3.4, p=0.15) (Table 3.6). In April 2010 the mean PIM 6 score for the children receiving NC was 17.8 (SD 4.8) and for the children receiving adapted NC in pairs was 17.3 (SD 4.2). The effect size was -0.54 (95% CI -1.17 to 0.10). The results demonstrate that children who received adapted NC in pairs scored lower, although not statistically significantly lower, on the PIM 6 mathematics test compared with children who had received NC individually (-2.3 95% CI -5.2 to 0.6, p=0.12). The pooling of the two sets of results shows no statistically significant difference between the scores of the children taught one-to-one or in pairs in terms of PIM 6 scores. However, there was a slight difference in favour of pairs.

- The secondary outcome measure was the Sandwell test (A or B depending on time of assessment). This measure was undertaken by people who knew whether the children were in the NC one-to-one group or the adapted NC pairs group. The effect size for the January 2010 testing was 0.11 (95% CI -0.46 to 0.68) and for the April 2010 testing was -0.84 (95% CI -1.46 to -0.22). However, we were unable to minimise the threat to the reliability of this measure due to the potential for a number of biases: the test was specifically developed to be the diagnostic part of NC; it is a narrow test, and it does not test broader mathematical abilities; the testers knew whether the children they were testing had received NC or adapted NC, and the testers were not independent.

3.1: Introduction

Although the previous trial found a statistically significant difference between normal classroom practice and Numbers Count (NC) teaching, a further question that is of importance is whether or not an adapted programme based on NC can be delivered to more than one child at the same time. Delivering a numeracy programme to pairs of children has obvious cost advantages over and above delivering a one-to-one programme. However, this cost saving may be gained at the expense of a less effective outcome for the child and may, therefore, not be cost-effective. In this chapter we answer the following question: is NC more or less effective than an adapted Numbers Count programme that can be delivered to two children at the same time?

3.2: Design and methods

3.2.1: Research objectives

Our primary aim was to obtain robust evidence of the *relative effectiveness* of the NC intervention (Numbers Count Handbook 2009-2010, Edge Hill University, 2008) when it is delivered mainly *individually to one child* and adapted NC when it is delivered to *groups of pairs of children* on attainment in mathematics. We undertook a pragmatic randomized controlled trial (RCT) in 15 schools. We compared one-to-one delivery of NC versus one to two (pairs) delivery of adapted NC. The outcome was attainment in mathematics. The Year 2 children identified by the 15 schools as being eligible to receive the intervention were randomized to receive an intervention individually or in groups of two during the school year 2009-10. All children eligible to receive an intervention received it. The 15 schools were selected using the following criteria:

- schools not participating in Trial 1;
- schools of sufficient size to enable more pupils to be identified for Numbers Count;
- schools in a local authority which had sufficient capacity to manage the implementation of a group work approach;
- Schools with an accredited Numbers Count teacher.

The wider impacts of the intervention were assessed by analysing wider quantitative outcomes of the children in the cohort (attitudes to mathematics, literacy and school).

Up to 14 eligible children in each of the recruited schools were identified by the school.

NC is mainly a one-to-one intervention[3]. For the purposes of this trial, Edge Hill gave permission for teachers to amend their approach in order to deliver an adapted version of the intervention to pairs of children, using NC resources. Teachers received some advice and support from Every Child a Chance Trust consultants on how this was to be done; however, it was up to individual teachers to amend their practice in order to meet the needs of pairs of children.

[3] As part of the NC lesson children may sometimes work in pairs, particularly towards the end of the programme.

There was a clear need to obtain reliable evidence to inform policy and practice, and crucially to establish the relative effectiveness of the mainly one-to-one delivery of Numbers Count compared with one-to-group delivery of adapted NC. The main focus of this trial arose from the recommendation in the Williams Independent Review of Mathematics Teaching in Early Years Settings and Primary Schools (Williams, 2008) that research should be conducted to establish whether individual or small group delivery of an intensive numeracy intervention is most effective and offers best value for money. A secondary focus was on a cost-effectiveness analysis of the various options (see Chapter 6).

This was a randomized controlled trial to assess both the relative effectiveness of NC intervention delivered one-to-one and adapted NC delivered in pairs. We also assessed the relative effectiveness of NC and adapted NC delivered in the autumn or the spring term. In this study the children within each school participating in the trial who were eligible to receive the intervention were randomly allocated a) to participate either individually or in groups, *and* b) to term of delivery. The participant schools for this trial were selected from the cohort of schools in which Numbers Count Teachers (NCTs) were implementing the NC for the *second* year in 2009-10 (excluding the schools where a new NCT was in training in 2009-10). We were able to assess the effectiveness of the intervention by using the data from children receiving the intervention individually comparing data from children receiving the intervention in pairs.

The protocol for Trial 2: Pairs (see 'Trial 2 Appendices' Appendix 1 in Torgerson et al, 2011c) emphasised the standardised training for delivery of NC and the standardised manual for implementation which are normal practice, but also justified the ways in which implementation of adapted NC was necessarily different from standardised practice *for the purposes of the trial.*

Schools identified the children who were eligible to receive an intervention, and consent was obtained from the children and their parents to be involved in the trial, (specifically to undertake any additional testing that was necessary for the purposes of the trial including consent to take the wider outcomes tests). Once consent had been checked and verified and the baseline testing had been completed, the schools contacted the Trial Co-ordinator either by telephone or by e-mail to access the randomization process, which was undertaken by the York Trials Unit. This ensured unbiased allocation to trial arm.

3.2.2: Avoidance of bias

We adopted the same methods to reduce the risk of bias we used in Trial 1: namely we used a robust random allocation method.

The design of this trial required 5 children in each school to receive NC or adapted NC in the autumn term 2009, and 5 children in each school to receive NC or adapted NC in the spring term 2010, and 2, 3 or 4 children to receive NC in the summer term 2010. Therefore 12, 13 or 14 eligible children in each of the recruited schools were identified by the school to receive an intervention. Each school had the flexibility to decide how many children would receive NC in the summer term (2, 3 or 4); this enabled the schools to keep 1 or 2 slots open to use for either teaching new children who arrived in the school during the year or for wider impact work within the school. Exceptionally schools could recommend a pupil as being unsuitable for randomization but this was discouraged as it would have reduced the external

validity of the trial. In the autumn term the Numbers Count Teacher (NCT) delivered NC to 1 child individually and adapted NC to 4 children in two pairs. In the spring term the NCT delivered NC 1 child individually and adapted NC to 4 children in two pairs. In the summer term NC was delivered to 2, 3 or 4 children individually. The University of York randomized the children to one-to-one delivery of NC or adapted NC. The teachers determined the makeup of the pairs, based on professional judgement, from the children randomly allocated to pairs. The University of York then randomly allocated the pairs to term of delivery.

Sample size and power – This trial had somewhat lower power than Trial 1. Given the smaller sample size in this study we did not have the same power to detect the relatively modest differences we might expect to occur between two 'active' interventions. The trial is reported as a pilot trial.

In this analysis we compared children who were randomized as individuals but were grouped in clusters (i.e., pairs). This grouping effect may have resulted in clustering of outcomes. Ignoring this clustering for the moment, our power calculation assumed the following: 0.70 correlation between pre and post-test; in 15 schools, a minimum of 30 children randomized to individual tuition and 120 randomized to pairs (4 pairs per school). For the sample size (i.e., 30 versus 120 children) we had approximately 80% power to show a difference of 0.55 of an effect size, assuming an intracluster correlation coefficient of 0.1 for the children in the pairs.

3.2.3: Economic evaluation

We undertook a cost effectiveness analysis comparing NC delivered to one child versus adapted NC delivered to pairs of children.

3.2.4: Outcome measures

As with Trial 1 the primary outcome measure was PIM 6. This was administered to all children and marked independently in January 2010 and in April 2010. Similarly all children received a pre-test in the form of the Sandwell test (A) at the beginning for the year. All children were post-tested at the end of the first term (Sandwell B) and again at the end of the second term (Sandwell A test). There was a final post-test at the end of the third term (Sandwell B test). The Sandwell test was the secondary outcome measure.

We also collected the KS1 mathematics results for all children selected for NC or adapted NC.

3.2.5: Wider impact

In addition to the assessment of impact on numeracy abilities, we measured the following variables in order to assess the wider impact of the intervention.

(a) Attention/behaviour/mental health (Strengths and Difficulties Questionnaire, SDQ Goodman (2001) teacher/parent scale) (sample of children, assessment not blinded);

(b) Attitudes to mathematics, literacy and school (Performance Indicators in Primary Schools, PIPS) (all children, assessment blinded).

All wider impact assessments were piloted before use (not in ECC schools), and administered independently (except for (a) SDQ Goodman which had to be conducted by a teacher who knew the child).

Table 3-1: Trial 2: Pairs Testing regime

Pupils	Baseline September 2009	January testing December 2009/January 2010	April testing March 2010/April 2010	July testing July 2010
5 children receiving NC or the adapted intervention in autumn term 2009	Sandwell A (Entry)	Sandwell B (Exit) INDEPENDENT TEST Progress in mathematics 6 Wider outcomes assessments (PIPS, SDQ)	Sandwell A (3 month)	Sandwell B (6 month)
5 children receiving NC or the adapted intervention in spring term 2010	Sandwell A	Sandwell B (Entry) INDEPENDENT TEST Progress in Mathematics 6 Wider outcomes assessment (PIPS, SDQ)	Sandwell A (Exit)	Sandwell B (3 month)
2,3 or 4 Children receiving NC in summer Term 2010	Sandwell A	Sandwell B INDEPENDENT TEST Progress in Mathematics 6 Wider Outcomes assessment (PIPS, SDQ)	Sandwell A (Entry)	Sandwell B (Exit)
Normal practice and required by evaluation				
Additional testing/assessment required by evaluation				

Table 3.2 gives the primary, secondary and exploratory outcomes, with timelines.

Table 3-2: Trial 2: Pairs Outcome measures

Primary outcome	Secondary outcomes	Exploratory outcomes
PIM 6 January (January testing) PIM 6 April (April testing)	Sandwell B December (January testing) Sandwell A March (April testing) Sandwell B July (July testing)	PIPS Quiz January (January testing) SDQ December (January testing) PIPS Quiz April (April testing) SDQ March (April testing)

A paper based survey was developed which sought factual information about teachers' experience and qualifications and included a log for the teachers to record each child's participation in NC. This information was used as a check for fidelity of implementation (but note this was by self-report, not independent observation, and conclusions derived from results take this limitation into account).

Table 3.3 gives the data collection information for the trial, with timelines.

Table 3-3: Trial 2: Pairs Data collection table

	PIM 6	Sandwell A	Sandwell B	PIPS Quiz	SDQ	Pupil Log	Teacher Survey
September 2009 (Baseline)		X					X
December 2009 (January testing)			X		X	X	
January 2010 (January testing)	X			X			
March 2010 (April testing)		X				X	
April 2010 (April testing)	X			X			
July 2010 (July testing)			X			X	X

3.2.6: Statistical analysis

All analyses were conducted on an intention to treat basis. Consequently any children who crossed over from either study arm were analysed as per their randomized allocation. Analyses were conducted in Stata using 2-sided significance tests at the 5% significance

level. All baseline data were summarised by treatment group and described. No formal statistical comparisons were undertaken for the baseline data. The primary outcome was the PIM 6. The scores on the PIM 6 were summarised descriptively (mean and standard deviation) by allocated group. Linear regression was used to compare the two groups with adjustments made for the potential clustering within schools using the Huber-White sandwich estimator (robust standard errors). The outcome modelled was the PIM 6 score and the model included age, gender, whether the child was receiving free school meals, Sandwell A test score (pre-test) and group allocation. This analysis was repeated for the secondary outcome which was the Sandwell test. (Note: the full results from the pupil logs and teacher surveys are presented in Appendices to the report; see Appendix 45 and Appendix 46, Trial 2 Appendices, Torgerson et al, 2011c).

The primary analysis (January testing) compared the children receiving NC individually in the autumn term with the children who were allocated to receive adapted NC in pairs in the autumn term. We also compared the children receiving NC individually in the autumn term with the children who were allocated to receive the NC intervention in the spring or summer terms; and similarly for children taught in pairs. The primary analysis (April testing) compared the children receiving NC individually in the spring term with the children who were allocated to receive adapted NC in pairs in the spring term. We also compared the children receiving NC individually in the spring term with the children who were allocated to receive NC in the summer term; and similarly for children taught in pairs.

The anonymity of all schools, children and teachers was preserved for all analyses and there is no presentation or comparison of the results from individual schools or teachers. Subgroup analyses assessed the effectiveness of the intervention for children with different learner characteristics (EAL, gender, free school meal status etc).

The results from the autumn and spring terms were combined, and subgroup analyses were undertaken to investigate the impact of term of delivery and the impact of training.

3.2.7: Quality assurance procedures for designing and reporting RCTs: the CONSORT guidelines

As with Trial 1 we designed, conducted and reported this study using the CONSORT guidelines.

3.2.8: Research ethics and data management

We submitted our research plans (protocol) for the trial to The University of York Humanities and Social Science Ethics Committee for ethical approval. Data processing and management abided by current data protection regulations. All data were stored on secure servers that are password protected. All electronic data can be held indefinitely. We used the SRA research ethics framework (see 'Trial 2 Appendices' Appendix 3 for full data protection issues, Torgerson et al, 2011c). We received approval for our protocol from DfE and the Steering Group. The trial protocol includes dates approval was received from the University of York HSSEC, the DfE and the Steering Group. All trial, ethics and testing protocols, information and consent forms, and all trial school correspondence templates are included in the Appendices to this document (Torgerson et al, 2011c).

3.3: Results

3.3.1: School progress through the trial

The progress of the schools through the trial is shown in Figure 3.1. Eleven schools were originally approached and invited to a recruitment conference. A further 30 schools were approached to consider taking part in a group work trial (pairs or triplets was not specified) and were invited to a recruitment conference. However, due to low recruitment numbers the decision was taken to include any schools that consented, from this additional 30 approached, in the Pairs trial. Consequently, in total 15 schools consented to take part in Trial 2: Pairs and these schools were randomized by the end of September 2009 and remained in the trial for the duration of the autumn term. One school had to withdraw from the trial at the end of the autumn term as the NC teacher was leaving and being replaced by a NC teacher who was not yet accredited. Fourteen schools remained in the trial as of January 2010. No schools withdrew from the trial during the spring term, thus 14 schools remained in the trial as of April 2010. One school withdrew at the beginning of the summer term as the NC teacher had to go on leave. Thus 13 schools remained in the trial at the end of the study.

3.3.2: Pupil progress through the trial

Each school was asked to identify between 12 and 14 children to take part in the trial.

- 9 schools selected 12 children
- 2 schools selected 13 children
- 4 schools selected 14 children

In total therefore 190 children selected by schools with parental/carer consent participated at the beginning of the trial. Three children from 1 school were excluded from being randomized to autumn term delivery of NC because they were also having Reading Recovery (RR) and the school did want the children to be having both RR and NC in the same term. These children were excluded from all analyses involving children allocated to autumn term.

At the beginning of the trial (after randomization) 1 child began receiving a behavioural intervention and therefore did not receive NC in Pairs in the autumn term (Group 2). The partner of this pair therefore received NC as an individual for the whole of the autumn term. During the autumn term 1 child left the trial because they moved to a new school. This child was receiving NC in a pair in the autumn term (Group 2). The partner of this pair therefore received NC as an individual for the remainder of the term. At the end of the autumn term 189 children remained in the trial.

At the beginning of the spring term 12 children were withdrawn because their school had withdrawn at the end of the autumn term. One child in Group 3 did not receive individual NC as their school felt the child was already performing well and would not benefit from the intervention; however data for this child were still collected and analysed. During the spring term an additional 2 children left the trial because they moved to another school (1 from

Group 3, 1 from Group 4). Thus at the end of the spring term 175 children remained in the trial.

At the beginning of the summer term 12 further children were withdrawn because their school withdrew from the trial at the beginning of the summer Term. At the beginning of the summer term 3 children left the trial because they moved school (1 from group 1, 1 from Group 4, 1 from Group 5). During the summer term an additional 2 children left the trial because they moved school (1 from Group 2, 1 from Group 4). At the end of the trial 158 children remained in the trial.

The progress of schools and children through the trial is shown in Figure 3.1

Figure 3-1: Trial 2 Pairs CONSORT diagram

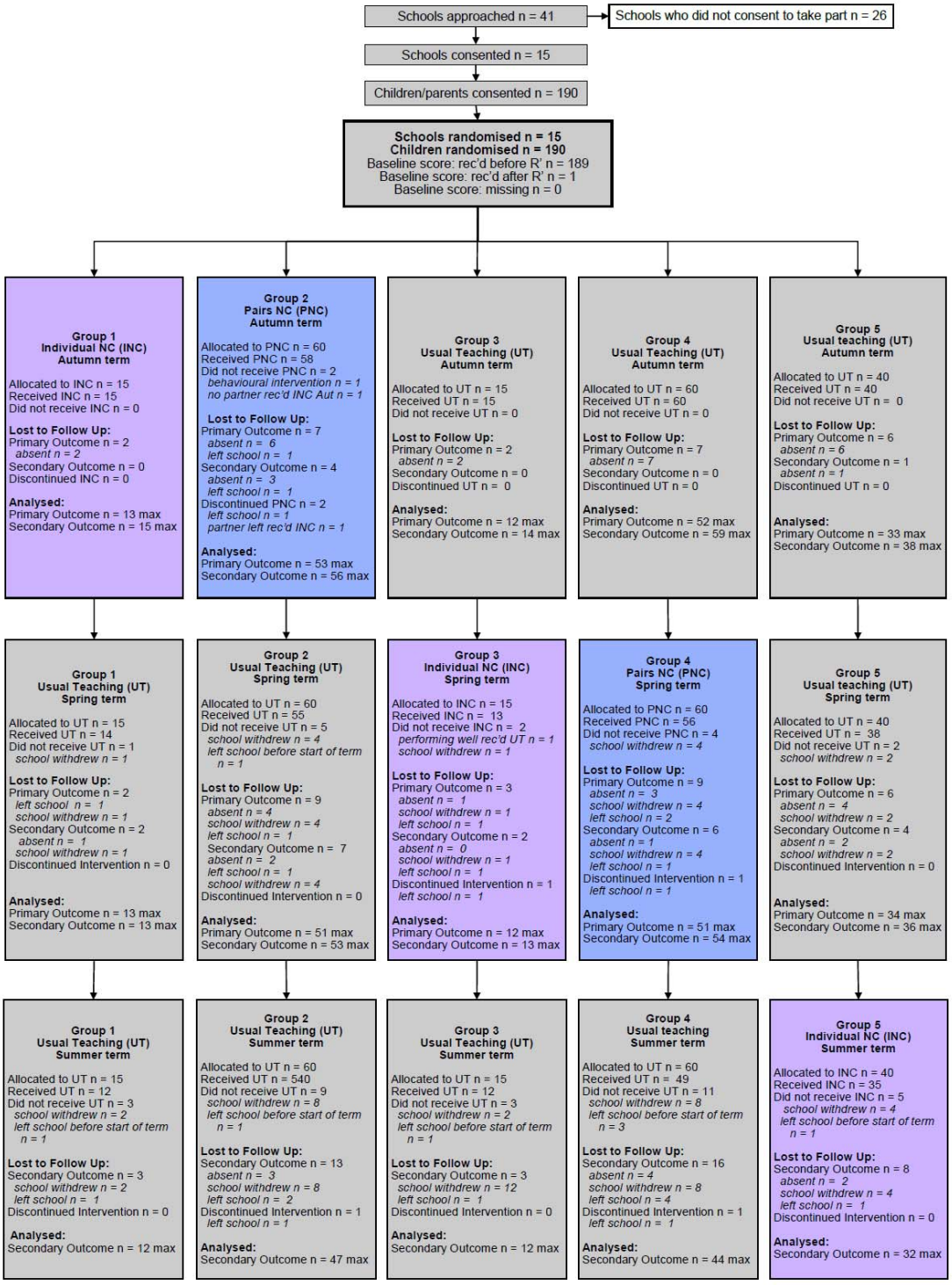


Table 3-4: Trial 2: Pairs Number of data points returned

	PIM 6	Sandwell A	Sandwell B	PIPS Quiz	SDQ Parent	SDQ Teacher	Pupil Log	Teacher Survey
September 2009 (Baseline)		190						15
December 2009 (January testing)			185		68	95	73	
January 2010 (January testing)	166			165				
March 2010 (April testing)		169			46	80	67	
April 2010 (April testing)	161			160				
July 2010 (July testing)			147				35	13

Maximum=190; for pupil log autumn max=75, spring max=75, summer max=40; for teacher survey max=15.

Protocol Deviations

- One child who was allocated to NC in a pair in autumn received NC individually in autumn as the partner needed to have a behavioural intervention.
- One child who was allocated to NC in a pair in the autumn term received NC individually in autumn as the partner left the trial to move to another school.
- One child who was allocated to NC in a pair in the autumn term was taught individually for part of a term, as the other child allocated to be in the pair left the school during the term.
- One child who was allocated to receive NC as an individual in the spring term was withdrawn from the intervention by their school as it was felt the intervention was no longer appropriate.
- Twelve children from 1 school were tested using the old version of the Sandwell test in September 2009.
- One child was tested late (after randomization) for the Sandwell A September 2009 testing.
- Eight children from 2 schools were tested late (after the Christmas holidays) for the Sandwell B December 2009 testing.
- Nine children from 1 school were tested using Sandwell B rather than Sandwell A at the March 2010 testing point.

- Six children from 5 schools were tested late (after the Easter holidays) for the Sandwell A March 2010 testing.

3.3.3: Baseline characteristics

Table 3.5 gives the baseline characteristics of the children included in this study. The characteristics are summarised by whether children were randomized to receive the NC intervention in the autumn, spring or summer term and by the method of delivery either one-to-one or one to two delivery. Three children were randomized to receive NC in the spring or summer term only and have been excluded from the summaries below.

Table 3-5: Trial 2: Pairs Baseline characteristics

Characteristic	Autumn		Spring		Summer
	Individual (N=15)	**Pairs (N=60)***	**Individual (N=14)**	**Pairs (N=59)**	**Individual (N=39)**
Age, mean (SD)	6.5 (0.3)	6.4 (0.3)	6.5 (0.3)	6.4 (0.3)	6.4 (0.3)
Sandwell A, mean (SD)	26.2 (5.9)	26.2 (7.8)	23.9 (8.7)	26.2 (6.7)	27.1 (6.6)
Free school meal, n (%)	7 (46.7)	31 (51.7)	9 (64.3)	29 (49.2)	18 (46.2)
Gender (females), n (%)	6 (40.0)	25 (41.7)	2 (14.3)	26 (44.1)	21 (53.9)

*Age was missing for one pupil (n=59)

We can see from Table 3.5 that the groups formed at baseline are comparable in age, Sandwell A mathematics scores, percentage of children receiving free school meals and gender.

3.3.4: Comparison of outcomes for one-to-one delivery and one-to-two delivery

Primary outcome

The primary outcome measure was the PIM 6 which was undertaken and marked blind to group allocation by independent testers.

Table 3-6: Trial 2: Pairs Primary outcome measure

Outcome	Individual	Pairs	Estimate*	ES
PIM 6 (January), mean (SD)**	15.5 (3.5) [n=13]	17.1 (5.0) [n=53]	1.4 (-0.6 to 3.4) [n=66]	0.30 (-0.31 to 0.91)
PIM 6 (April), mean (SD)	17.8 (4.8) [n=12]	17.3 (4.2) [n=51]	-2.3 (-5.2 to 0.6) [n=63]	-0.54 (-1.17 to 0.10)

* Analyses were adjusted for baseline Sandwell A test scores, age, gender, free school meals and the clustering within schools.
** Analysis excludes children unable to be randomized to autumn term.

January testing

The mean PIM 6 score for the children receiving the NC intervention individually was 15.5 (SD 3.5) and for the children receiving adapted NC in pairs was 17.1 (SD 5.0). The effect size was 0.30 (95% CI -0.31 to 0.91). The results demonstrate that children who received adapted NC in pairs scored slightly higher, although not statistically significantly higher, on the PIM 6 mathematics test compared with children who had received NC individually (1.4 95% CI -0.6 to 3.4, p=0.15) (Table 3.6).

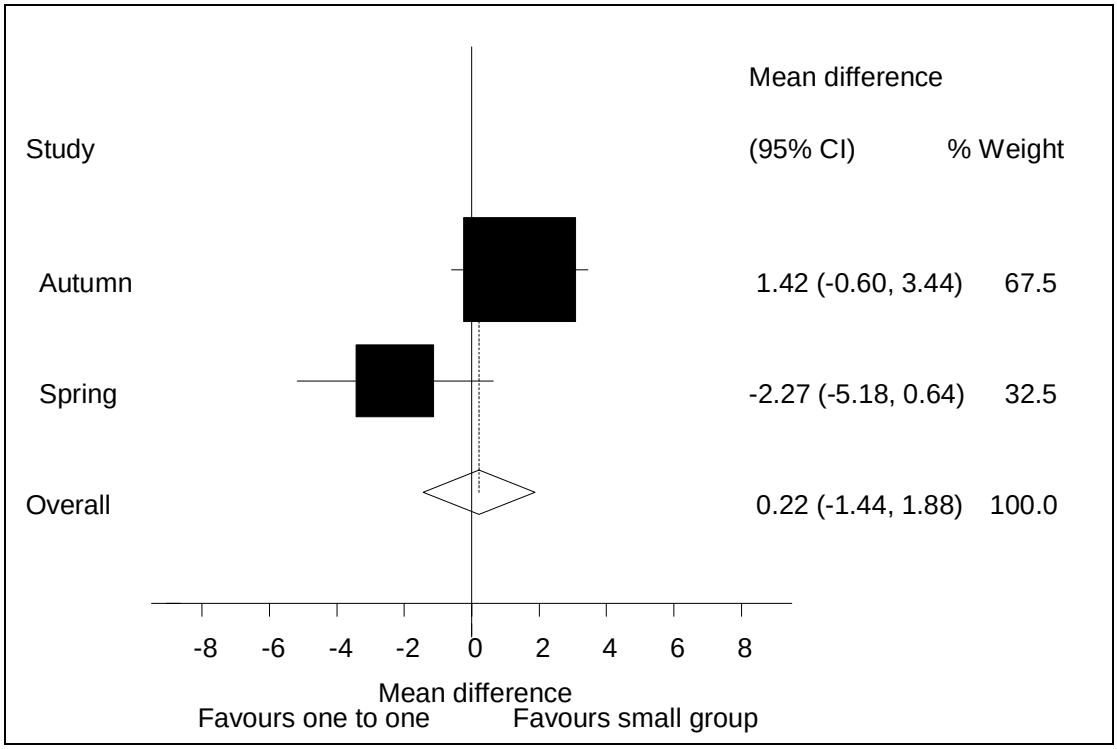
April testing

The mean PIM 6 score for the children receiving the NC intervention individually was 17.8 (SD 4.8) and for the children receiving adapted NC in pairs was 17.3 (SD 4.2). The effect size was -0.54 (95% CI -1.17 to 0.10). The results demonstrate that children who received adapted NC in pairs scored lower, although not statistically significantly lower, on the PIM 6 mathematics test compared with children who had received the NC intervention individually (-2.3 95% CI -5.2 to 0.6, p=0.12) (Table 3.6).

Overall

In figure 3.2 we combine the two analyses of one-to-one teaching versus teaching in pairs. As the figure demonstrates the pooling of the two trials shows no statistically significant difference between the scores of the children taught individually or in pairs in terms of PIM 6 scores although a slight difference in favour of pairs was observed.

Figure 3-2: Forest plot of NC one-to-one vs. one-to-two teaching



In Figure 3-2 we show graphically the results of the autumn and spring results. There are two study plots: the first shows the estimated effect of the study with horizontal lines showing how uncertain we are about the results. There is uncertainty about the results because of the relatively small sample size and so we can see that the horizontal lines cross over the vertical axis which is zero. In other words no statistically significant effect was observed. The second plot is on the other side of no effect but its horizontal lines also pass through zero or the no effect vertical line. Finally the two studies are combined in the diamond figure at the bottom of the graph. The peak of the diamond is just past the right hand side of the line of zero indicating a small benefit in favour of one-to-two delivery, which is not statistically significant because part of the diamond passes over the vertical line of no effect.

Secondary outcome

The secondary outcome measure was the Sandwell test (A or B depending on time of assessment). This measure was not undertaken blind or marked blind to group allocation.

Table 3-7: Trial 2: Pairs Secondary outcomes

Outcome	Individual	Pairs	Estimate*	ES
Sandwell B (January), mean (SD)**	43.4 (12.2) [n=15]	44.7 (13.3) [n=56]	1.4 (-1.9 to 4.7) [n=71]	0.11 (-0.46 to 0.68)
Sandwell A (April), mean (SD)	56.5 (12.9) [n=13]	48.5 (11.4) [n=54]	-10.8 (-19.3 to -.24) [n=67]	-0.84 (-1.46 to -0.22)

* Analyses were adjusted for baseline Sandwell A test scores, age, gender, free school meals and the clustering within schools.

**Analysis excludes children unable to be randomized to autumn term.

The effect size for the January testing was 0.11 (95% CI -0.46 to 0.68) and for the April testing was -0.84 (95% CI -1.46 to -0.22). However, we were unable to minimise the threat to internal reliability of this measure due to the potential for outcome ascertainment bias.

3.3.5: Comparison of outcomes by term of delivery

We were also able to compare the first cohort to the second cohort on the April assessment to compare delivery of the intervention in autumn and spring terms and all three cohorts on the July assessment to see if there was any difference in effectiveness depending on term of delivery.

Primary outcome

The primary outcome measure was the PIM 6 which was undertaken and marked blind to group allocation by independent testers.

Table 3-8: Trial 2: Pairs Primary outcome measure for delivery of intervention

Outcome	Autumn	Spring	Estimate*	ES
One-to-one, mean (SD)	16.3 (5.0) [n=13]	17.8 (4.8) [n=12]	2.5 (-1.8 to 6.9) [n=25]	0.5 (-0.3 to 1.3)
One-to-two, mean (SD)	17.9 (4.3) [n=51]	17.3 (4.2) [n=51]	-1.0 (-2.3 to 0.2) [n=102]	-0.2 (-0.6 to 0.2)

* Analyses were adjusted for baseline Sandwell A test scores, age, gender, free school meals and the clustering within schools. Analyses exclude children who could not be randomized to autumn term.

One-to-one summary

The mean PIM 6 score for the children receiving NC individually in the autumn term was 16.3 (SD 5.0) and for the children receiving NC individually in the spring term was 17.8 (SD 4.8). The results demonstrate that children who received NC individually in autumn scored slightly lower, although not significantly lower, on the PIM 6 mathematics test compared to children who received NC individually in spring (2.5 95% CI -1.8 to 6.9, p=0.23) (Table 3.8).

One to two summary

The mean PIM 6 score for the children receiving adapted NC in pairs in the autumn term was 17.9 (SD 4.3) and for children receiving adapted NC in pairs in the spring term was 17.3 (SD 4.2). The results demonstrate that children who received adapted NC in pairs in autumn scored slightly higher, although not significantly higher, on the PIM 6 mathematics test compared to children who received adapted NC intervention in pairs in spring (-1.0 95% CI -2.3 to 0.2, p=0.1) (Table 3.8).

Secondary outcome

The secondary outcome measure was the Sandwell test (A or B depending on time of assessment). This measure was not undertaken blind or marked blind to group allocation. For the secondary outcome measure we were able to compare the mean differences of one-

to-one delivery in autumn term to one-to-one delivery in spring term and one-to-two delivery in autumn term to one-to-two delivery in spring term, both on the April assessment.

Table 3-9: Trial 2: Pairs Secondary outcome measure for delivery of intervention

Outcome	Autumn	Spring	Estimate*	ES
One-to-one, mean (SD)	46.9 (13.2) [n=13]	56.5 (12.9) [n=13]	9.7 (-1.6 to 21.1) [n=26]	0.8 (-0.02 to 1.6)
One-to-two, mean (SD)	49.5 (12.6) [n=53]	48.5 (11.4) [n=54]	-2.1 (-6.0 to 1.7) [[n=107]	-0.2 (-0.6 to 0.2)

* Analyses were adjusted for baseline Sandwell A test scores, age, gender, free school meals and the clustering within schools. Analyses exclude children who could not be randomized to autumn term.

The effect size for the one-to-one comparison was 0.8 (95% CI -0.02 to 1.6) and for the one-to-two comparison was -0.2 (95% CI -0.6 to 0.2). However, we were unable to minimise the threat to internal reliability of this measure due to the potential for outcome ascertainment bias. We were also able to compare one-to-one delivery across the three terms on the July assessment. The results of this analysis are presented in Table 3.10.

Table 3-10: Trial 2: Pairs Secondary outcome scores

Outcome	Autumn	Spring	Summer	p-value*
Sandwell B (July), mean (SD)	50.8 (14.9) [n=12]	58.3 (13.3) [n=12]	56.0 (11.5) [n=32]	0.16

*Analysis excludes children who could not be randomized to autumn term [n=56].

The mean Sandwell B score for the children receiving NC individually in the autumn term was 50.8 (SD 14.9), in the spring term was 58.3 (SD 13.3) and the summer term was 56.0 (11.5). The results demonstrate that there was little or no evidence of an overall difference in mean Sandwell B scores among the three terms (p=0.16; Table 3.10).

3.3.6: Comparison of outcomes for one-to-one delivery and waiting list control

We were also able to compare the mean differences of one-to-one intervention children versus waiting list control children for the January and April testing. This is the same analysis as the analysis undertaken in Trial 1 which investigated the short-term impact of the intervention group receiving NC compared with the control group not receiving NC. Therefore, these data can be used to supplement the results from Trial 1.

Table 3-11: Trial 2: Pairs one-to-one delivery vs. waiting list control

Outcome	Intervention	Control	Estimate*	ES
PIM 6 (January), mean (SD)**	15.5 (3.5) [n=13]	15.2 (4.5) [n=100]	0.7 (-0.8 to 2.3) [n=113]	0.2 (-0.4 to 0.7)
PIM 6 (April), mean (SD)	17.8 (4.8) [n=12]	17.6 (4.3) [n=34]	2.2 (-0.2 to 4.6) [n=46]	-0.5 (-0.2 to 1.2)
Sandwell B (January), mean (SD)**	43.4 (12.2) [n=15]	31.7 (9.2) [n=111]	11.5 (6.1 to 17.0) [n=126]	1.3 (0.7 to 1.8)
Sandwell A (April), mean (SD)	56.5 (12.9) [n=13]	38.7 (9.7) [n=36]	19.0 (11.6 to 26.4) [n=49]	2.0 (1.2 to 2.7)

* Analyses were adjusted for baseline Sandwell A test scores, age, gender, free school meals and the clustering within schools.
**Analyses exclude children who could not be randomized to autumn term.

January testing

The mean PIM 6 score for the children receiving NC individually in the autumn term was 15.5 (SD 3.5) and for the children who had yet to receive the intervention was 15.2 (SD 4.5). The results demonstrate that children who received NC scored slightly higher, although not significantly higher, on the PIM 6 mathematics test compared to children who had yet to receive the NC intervention (0.7 95% CI -0.8 to 2.3, p=0.33) (Table 3.11).

April testing

The mean PIM 6 score for the children receiving NC individually in the spring term was 17.8 (SD 4.8) and for the children who had yet to receive the intervention was 17.6 (SD 4.3). The results demonstrate that children who received NC scored higher, although not significantly higher, on the PIM 6 mathematics test compared to children who had yet to receive NC (2.2 95% CI -0.2 to 4.6, p=0.07) (Table 3.11).

3.3.7: Comparison of outcomes for one to two delivery and waiting list control

We were also able to compare the mean differences of one to two intervention children versus waiting list control children for the January and April testing. This is the same analysis as the analysis undertaken in Trial 1 which investigated the short-term impact of the intervention group receiving NC compared with the control group not receiving NC. Therefore, these data can be used to supplement the results from Trial 1.

Table 3-12: Trial 2: Pairs one to two delivery vs. waiting list control

Outcome	Intervention	Control	Estimate*	ES
PIM 6 (January), mean (SD)**	17.1 (5.0) [n=53]	15.2 (4.5) [n=100]	2.3 (1.0 to 3.5) [n=153]	0.5 (0.16 to 0.84)
PIM 6 (April), mean (SD)	17.3 (4.2) [n=51]	17.6 (4.3) [n=34]	0.1 (-1.7 to 1.9) [n=85]	0.01 (-0.42 to 0.45)
Sandwell B (January), mean (SD)**	44.7 (13.3) [n=56]	31.7 (9.2) [n=111]	13.0 (9.6 to 16.3) [n=167]	1.4 (1.1 to 1.8)
Sandwell A (April), mean (SD)	48.5 (11.4) [n=54]	38.7 (9.7) [n=36]	10.1 (6.9 to 13.3) [n=90]	1.1 (0.6 to 1.5)

* Analyses were adjusted for baseline Sandwell A test scores, age, gender, free school meals and the clustering within schools.
**Analyses exclude children who could not be randomized to autumn term.

January testing

The mean PIM 6 score for the children receiving adapted NC in pairs in the autumn term was 17.1 (SD 5.0) and for children who had yet to receive the intervention was 15.2 (SD 4.5). The effect size was 0.5 (95% CI 0.16 to 0.84) indicating strong evidence of a difference between the two groups (2.3 95% CI 1.0 to 3.5, p=0.002) (Table 3.12). The results demonstrate that the children who received adapted NC in pairs in the autumn term had higher scores on the PIM 6 mathematics test compared to the children who had not yet received NC.

April testing

The mean PIM 6 score for the children receiving adapted NC in pairs in the spring term was 17.3 (SD 4.2) and for children who had yet to receive the intervention was 17.6 (SD 4.3). The effect size was 0.01 (95% CI -0.42 to 0.45) indicating little or no evidence of a difference between the two groups (0.1 95% CI -1.7 to 1.9, p=0.94) (Table 3.12). The results demonstrate that the children who received adapted NC in pairs in the spring term had similar scores on the PIM 6 mathematics test compared to the children who had not yet received NC.

3.4: Discussion

In Trial 2 we compared outcomes for Numbers Count (NC) delivered as a one-to-one intervention with adapted NC delivered to pairs of children or to normal classroom practice. The numbers in this evaluation are relatively small compared to the numbers in Trial 1. In the January testing we found a difference, which was not statistically significant, favouring

children being taught adapted NC in pairs. However, in the April testing we found a non-statistically significant difference favouring children taught individually. When we combined these two sets of results in a meta-analysis we found a slight difference favouring the teaching of adapted NC to children in pairs, but this difference was not statistically significant.

The confidence intervals around the estimate of effect between the two modes of delivery (one-to-one and pair) were very wide; therefore we cannot say which is the more effective mode of teaching. A larger suitably powered randomized controlled trial evaluating the relative effectiveness of one-to-one teaching (NC) and pair (adapted NC) would be required to answer this important question.

This study has a number of strengths and limitations. The key strengths are similar to those of Trial 1: namely, a robust randomization procedure and independent testing. However, in contrast to Trial 1, we had relatively few participants in Trial 2 leading to relatively wide confidence intervals around the estimates of effect.

It is appropriate to note that that schools volunteered to take part in the pair (and triplets trial, see Chapter 4), as all schools had volunteered to take part in ECC, and this could have created a biased sample in any individual versus pairs comparisons. However, from the process evaluation site visits (which were light touch for pairs) we did not see any evidence which led us to believe that Trial 2: Pairs schools were significantly or systematically different from Trial 1 schools.

3.5: Conclusions

In summary, our data, within the limitations of the small sample size, suggest that adapted NC delivered in pairs is similar in effectiveness to one-to-one delivery of NC.

Chapter 4: Impact: Trial 2: Triplets

Key summary points

- We undertook a feasibility randomized controlled trial evaluating the effectiveness of Numbers Count compared with adapted Numbers Count delivered to children in triplets for attainment in mathematics.
- The primary analyses are based on data from 45 children (January 2010 testing) and 47 children (April 2010 testing).
- There were no statistically significant differences in outcome comparing children receiving Numbers Count and children receiving adapted Numbers Count delivered in triplets.

4.1: Introduction

The previous trials indicated the effectiveness of Numbers Count (NC) compared with normal classroom teaching and adapted Numbers Count delivered to children in pairs. In the feasibility trial described in this chapter we compared adapted Numbers Count delivered to three children at a time. We were trying to answer the question: Is adapted Numbers Count delivered to children in triplets more or less effective than NC?

4.2: Design and methods

We undertook a pragmatic feasibility randomized controlled trial evaluating the effectiveness of Numbers Count delivered individually versus group delivery of adapted NC for attainment in mathematics of the children in Year 2, selected by participating schools, for their low performance in mathematics. This was a focused randomized controlled trial to assess the relative effectiveness of NC delivered one-to-one and adapted NC in triplets. We also assessed the relative effectiveness of NC and adapted NC delivered in the autumn or the spring term.

In Trial 2: Triplets, the children within each school participating in the trial who were eligible to receive the intervention were randomly allocated a) to participate either individually or in groups of three children, *and* b) to term of delivery. The participant schools for this trial were selected from the cohort of schools in which Numbers Count Teachers (NCTs) were implementing the intervention for the *second* year in 2009-10 (excluding the schools where a new NC teacher was in training in 2009-10). We were able to assess the effectiveness of the intervention by using the data from children receiving the intervention individually compared with data from children receiving the intervention in groups of three children.

Our protocol for this trial emphasised the standardised training for delivery of NC and the standardised manual for implementation of the Numbers Count intervention which are normal practice, but also justified the ways in which implementation of the adapted intervention was necessarily different from standardised practice *for the purposes of the trial.*

Schools identified the children who were eligible to receive an intervention, and consent was obtained from the children and their parents to be involved in the trial, (specifically to

undertake any additional testing that was necessary for the purposes of the trial including consent to take the wider outcomes tests). Once consent had been checked and verified and the baseline testing had been completed, the schools contacted the Trial Co-ordinator either by telephone or by e-mail to access the randomization process, which was undertaken by the York Trials Unit and which ensured unbiased allocation to trial arm.

4.2.1: Avoidance of bias

As with the previous trials we used a secure randomization system, to avoid selection bias.

The design of this trial required 7 children in each school to receive NC or adapted NC in the autumn term 2009, 7 children in each school to receive NC or adapted NC in the spring term 2010, and 2, 3 or 4 children to receive NC in the summer term 2010. Therefore 16, 17 or 18 eligible children in each of the recruited schools were identified by the school to receive an intervention. Each school had the flexibility to decide how many children would receive NC in the summer term (2, 3 or 4) which enabled the schools to keep 1 or 2 slots open to use for either teaching new children who arrived in the school during the year or for wider impact work within the school. Exceptionally schools could recommend a pupil as being unsuitable for randomization, but this was discouraged as it would have reduced the external validity of the trial. In the autumn term the NC teacher in each school delivered NC to 1 child individually and adapted NC to 6 children in two triplets. In the spring term the NC teacher delivered NC to 1 child individually and adapted NC to 6 children in two triplets. In the summer term NC was delivered to 2, 3 or 4 children individually. The teachers determined the makeup of the triplets, based on professional judgement, from the children randomly allocated to delivery of adapted NC in triplets.

[Note: Barclays agreed to fund an additional 4 slots for 4 children to receive Numbers Count at some schools. In order that this would not introduce a potential source of bias (selection bias) we asked the schools affected to nominate an additional 4 children, and these were included in the randomization. However, once children had been randomly allocated to be 'children funded by Barclays' they were not included in the trial. Please see 'Trial 2 Appendices' Appendix 1: Appendix C in Torgerson et al, 2011c for the trial design diagram for schools with Barclays funding.

Sample size and power – The anticipated sample size in the trial did not give us sufficient statistical power to be confident of our findings. Consequently this study can only be regarded as a feasibility study.

4.2.2: Outcome measures

The primary outcome measure is the PIM 6 test. This was administered and marked independently in January 2010, and in April 2010.

All children selected for NC or adapted NC received a pre-test in the form of the Sandwell test A at the beginning for the year. All children were post-tested at the end of the first term (Sandwell B) and again at the end of the second term (Sandwell A test). A final post-test was conducted at the end of the third term (Sandwell B test). The Sandwell test was the secondary outcome measure.

4.2.3: Wider impact

In addition to the assessment of impact on numeracy abilities, we measured the following variables in order to assess the wider impact of the intervention:

(a) Attention/behaviour/mental health (Strengths and Difficulties Questionnaire, SDQ Goodman teacher/parent scale);

(b) Attitudes to mathematics, literacy and school (Performance Indicators in Primary Schools, PIPS).

All wider impact assessments were piloted before use (not in ECC schools), and administered independently, except for (a) SDQ Goodman which needed to be conducted by a teacher who knew the child.

Table 4-1: Trial 2: Triplets Testing regime

<table>
<tr><th>Pupils</th><th>Baseline September 2009</th><th>January testing December 2009/January 2010</th><th>April testing March 2010/April 2010</th><th>July testing July 2010</th></tr>
<tr><td rowspan="3">7 children receiving NC or adapted NC in autumn term 2009</td><td rowspan="3">Sandwell A (Entry)</td><td>Sandwell B (Exit)</td><td rowspan="3">Sandwell A (3 month)</td><td rowspan="3">Sandwell B (6 month)</td></tr>
<tr><td>INDEPENDENT TEST Progress in mathematics 6</td></tr>
<tr><td>Wider outcomes assessments (PIPS, SDQ)</td></tr>
<tr><td rowspan="3">7 children receiving NC or adapted NC in spring term 2010</td><td rowspan="3">Sandwell A</td><td>Sandwell B (Entry)</td><td rowspan="3">Sandwell A (Exit)</td><td rowspan="3">Sandwell B (3 month)</td></tr>
<tr><td>INDEPENDENT TEST Progress in Mathematics 6</td></tr>
<tr><td>Wider outcomes assessment (PIPS, SDQ)</td></tr>
<tr><td rowspan="3">2,3 or 4 children receiving NC in summer term 2010</td><td rowspan="3">Sandwell A</td><td>Sandwell B</td><td rowspan="3">Sandwell A (Entry)</td><td rowspan="3">Sandwell B (Exit)</td></tr>
<tr><td>INDEPENDENT TEST Progress in Mathematics 6</td></tr>
<tr><td>Wider Outcomes assessment (PIPS, SDQ)</td></tr>
<tr><td colspan="5">Normal practice and required by evaluation</td></tr>
<tr><td colspan="5">Additional testing/assessment required by evaluation</td></tr>
</table>

Table 4-2: Trial 2: Triplets Outcome measures

Primary outcome	Secondary outcomes	Exploratory outcomes
PIM 6 January (January testing) PIM 6 April (April testing)	Sandwell B December (January testing) Sandwell A March (April testing) Sandwell B July (July testing)	PIPS Quiz January (January testing) SDQ December (January testing) PIPS Quiz April (April testing) SDQ March (April testing)

Table 4.2 gives the primary, secondary and exploratory outcomes for the trial, with timelines.

Table 4-3: Trial 2: Triplets Data collection table

	PIM 6	Sandwell A	Sandwell B	PIPS Quiz	SDQ	Pupil Log	Teacher Survey
September 2009 (Baseline)		X					X
December 2009 (January testing)			X		X	X	
January 2010 (January testing)	X			X			
March 2010 (April testing)		X				X	
April 2010 (April testing)	X			X			
July 2010 (July testing)			X			X	X

Table 4.3 gives the data collection framework, with timelines

4.2.4: Statistical analysis

The anonymity of all schools, children and teachers was preserved for all analyses, and there is no presentation or comparison of the results from individual schools or teachers.

The results from the autumn and spring terms were combined.

4.2.5: Quality assurance procedures for designing and reporting RCTs: the CONSORT guidelines

As with the two previous trials we designed, conducted and reported the results of this trial in line with the CONSORT guidelines.

4.2.6: Research ethics and database management

We submitted our research plan (protocol) for the trial to The University of York Humanities and Social Science Ethics Committee for ethical approval. Data processing and management abided by current data protection regulations. All data were stored on secure servers that are password protected. We used the SRA research ethics framework. We received approval for our protocol from DfE and the Steering Group. The trial protocol includes dates approval was received from the University of York HSSEC, the DfE and the Steering Group. All trial, ethics and testing protocols, information and consent forms, and all trial school correspondence templates are included in the Appendices to this document (Torgerson et al, 2011c).

4.3: Results

4.3.1: School progress through the trial

Data from 8 schools were included in the final analysis. 17 schools were approached to take part in Trial 2: Triplets and were invited to a recruitment conference Eight schools consented to take part. Pupils in all 8 schools were randomized by the end of September 2009 and remained in the trial for the duration of the study.

4.3.2: Pupil progress through the trial

Each school was asked to identify between 16 and 18 children to take part in the trial if they did not have any extra funding for Barclays (1 school) and between 20 and 22 if they did (7 schools). One school selected 17 children. In this school, 2 children were randomized to individual delivery of NC in the autumn, 3 to be taught in one triplet in the autumn term, 2 children to individual delivery of NC in the spring term, 3 to be taught in one triplet in the spring term, 3 to individual delivery of NC in the summer term and 4 to be funded by Barclays.

- 1 school selected 17 children (as above)
- 1 school selected 18 children
- 1 school selected 19 children
- 4 schools selected 20 children
- 1 school selected 22 children

In total, 156 children selected by schools with parental consent were included in the trial. Twenty seven children were randomly allocated as ‘funded by Barclays’ and from this point on were not considered as part of the trial, leaving 129 children in the trial. One child was excluded from being randomized to autumn term delivery, because he/she was also having Reading Recovery (RR) and the school did not want him/her to have both RR and NC in the

same term. This child was excluded from all analyses involving children allocated to autumn term.

During the autumn term 2 children left the trial because they moved to a new school (1 from Group 2, 1 from Group 3). The child in Group 2 was receiving an adapted intervention in a triplet. The two remaining children in their triplet were taught as a pair for the remainder of the term. At the end of the autumn term 127 children remained in the trial.

At the beginning of the spring term 1 child from Group 3 left the trial because they moved school. During the spring term 2 additional children left the trial because they moved school (1 from Group 4, 1 from Group 5). The child in Group 4 was receiving an adapted intervention in a triplet. The two remaining children in their triplet were taught as a pair for the remainder of the term. Thus, at the end of the spring term 124 children remained in the trial.

At the beginning of the summer term 2 children left the trial because they moved school (1 from Group 2, 1 from Group 5). During the summer term 1 further child from group 1 left the trial because they moved school. At the end of the trial 121 children remained in the trial.

Figure 4.1 demonstrates the progress of the schools and children through the trial.

Figure 4-1: Trial 2: Triplets CONSORT diagram

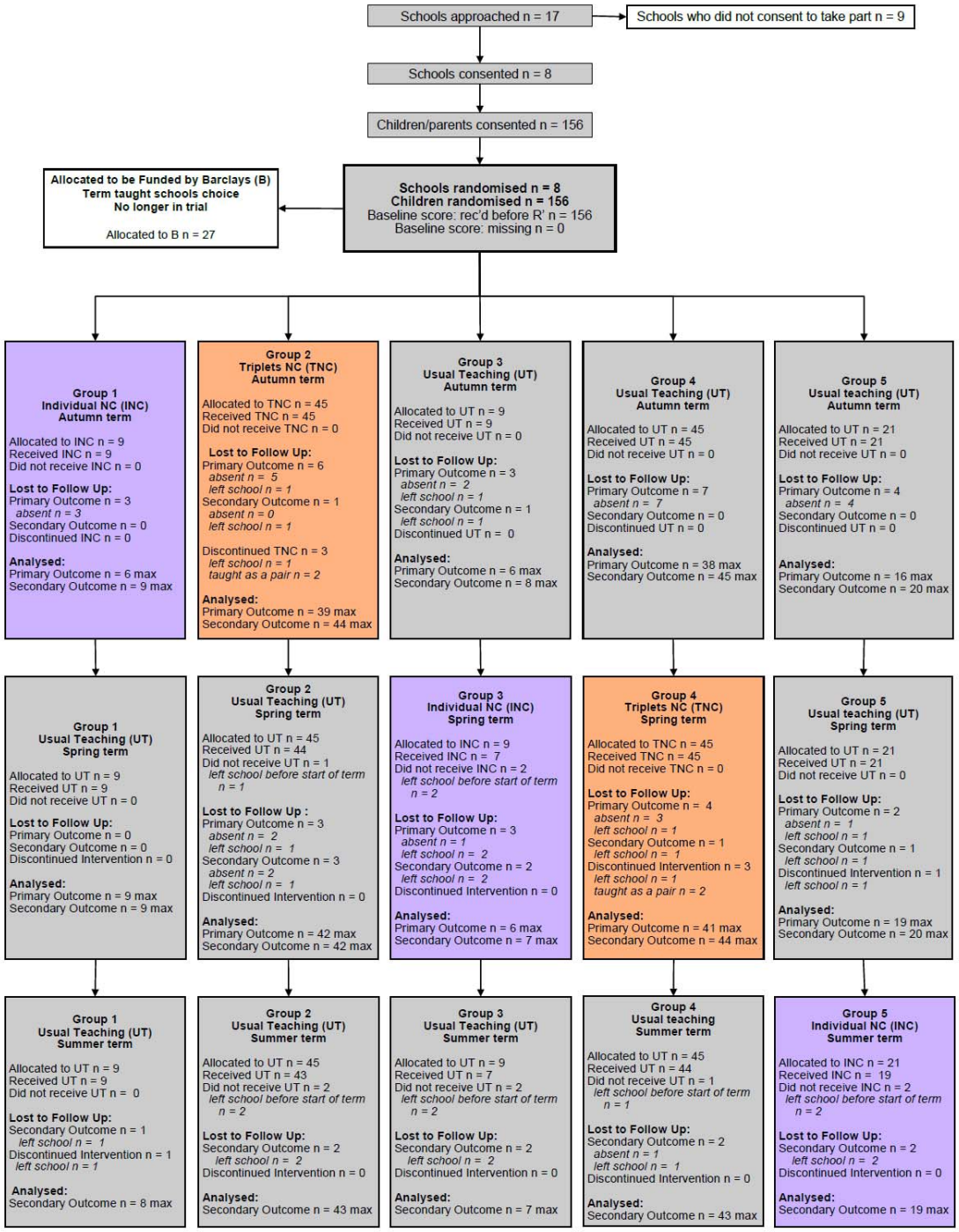


Table 4-4: Trial 2: Triplets Number of data points returned

	PIM 6	Sandwell A	Sandwell B	PIPS Quiz	SDQ Parent	SDQ Teacher	Pupil Log	Teacher Survey
September 2009 (Baseline)		129						8
December 2009 (January testing)			127		24	75	53	
January 2010 (January testing)	106			71				
March 2010 (April testing)		122			13	85	52	
April 2010 (April testing)	117			117				
July 2010 (July testing)			120				14	5

Maximum=156; for pupil log autumn max=54, spring max=54, summer max=21; for teacher survey max=8.

Protocol Deviations

- Four children who were allocated to be taught in a triplet, were actually taught in a pair for part of a term, as two triplet groups each had one child who left the school during the term, leaving the remaining two children to be taught in a pair.
- Sixteen children from one school were tested using the old version of the Sandwell test in September 2009 (rather than the Sandwell A test) and December 2009 (rather than the Sandwell B test).
- Six children from 3 schools were tested late (after the Easter holidays) for the Sandwell A March 2010 testing.

4.3.3: Baseline characteristics

Table 4.5 gives the baseline characteristics of the children included in this study. The characteristics are summarised by whether children were randomized to receive NC in the autumn, spring or summer term and by the method of delivery (either one-to-one or one to three). One child could not be allocated to autumn term and has been excluded from the summaries presented below.

Table 4-5: Trial 2: Triplets Baseline characteristics

Characteristic	Autumn		Spring		Summer
	Individual (N=9)	Triplets (N=45)[1]	Individual (N=9)[2]	Triplets (N=45)[3]	Individual (N=20)
Age, mean (SD)	6.5 (0.2)	6.4 (0.2)	6.5 (0.3)	6.5 (0.3)	6.5 (0.3)
Sandwell A, mean (SD)	32.6 (5.1)	33.2 (6.9)	32.1 (5.5)	29.1 (6.3)	28.9 (8.7)
Free school meal, n (%)	2 (22.2)	13 (28.9)	3 (37.5)	23 (51.1)	6 (30.0)
Gender (females), n (%)	7 (77.8)	22 (50.0)	3 (33.3)	17 (37.8)	10 (50.0)

[1]Gender was missing for one pupil (n=44) [2]Age was missing for one pupil (n=8) and free school meal status was missing for one pupil (n=8) [3]Free school meal status was missing for one pupil (n=44)

We can see from Table 4.5 that the groups formed at baseline were comparable in age, Sandwell A mathematics scores, percentage of children receiving free school meals and gender.

4.3.4: Comparison of outcomes for one-to-one delivery and one-to-three delivery

Primary outcome

The primary outcome measure was the PIM 6 which was undertaken and marked blind to group allocation by independent testers.

Table 4-6: Trial 2: Triplets Primary outcome measure

Outcome	Individual	Triplets	Estimate*	ES
PIM 6 (January), mean (SD)**	17.5 (4.3) [n=6]	18.0 (4.4) [n=39]	1.0 (-2.4 to 4.4) [n=45]	0.23 (-0.63 to 1.09)
PIM 6 (April), mean (SD)	22.8 (3.4) [n=6]	19.2 (5.4) [n=41]	-1.6 (-4.6 to 1.3) [n=47]	-0.32 (-1.18 to 0.54)

* Analyses were adjusted for baseline Sandwell A test scores, age, gender, free school meals and the clustering within schools

** Analysis excludes children who could not be randomized to autumn term.

January testing

The mean PIM 6 score for the children receiving NC individually was 17.5 (SD 4.3) and for the children receiving adapted NC in triplets was 18.0 (SD 4.4). The results highlight the finding that children who received adapted NC in triplets scored slightly higher, although not statistically significantly higher, on the PIM 6 mathematics test compared with children who had received NC individually (1.0 95% CI -2.4 to 4.4, p=0.51). The effect size was 0.23 (95% CI -0.63 to 1.09) (Table 4.6).

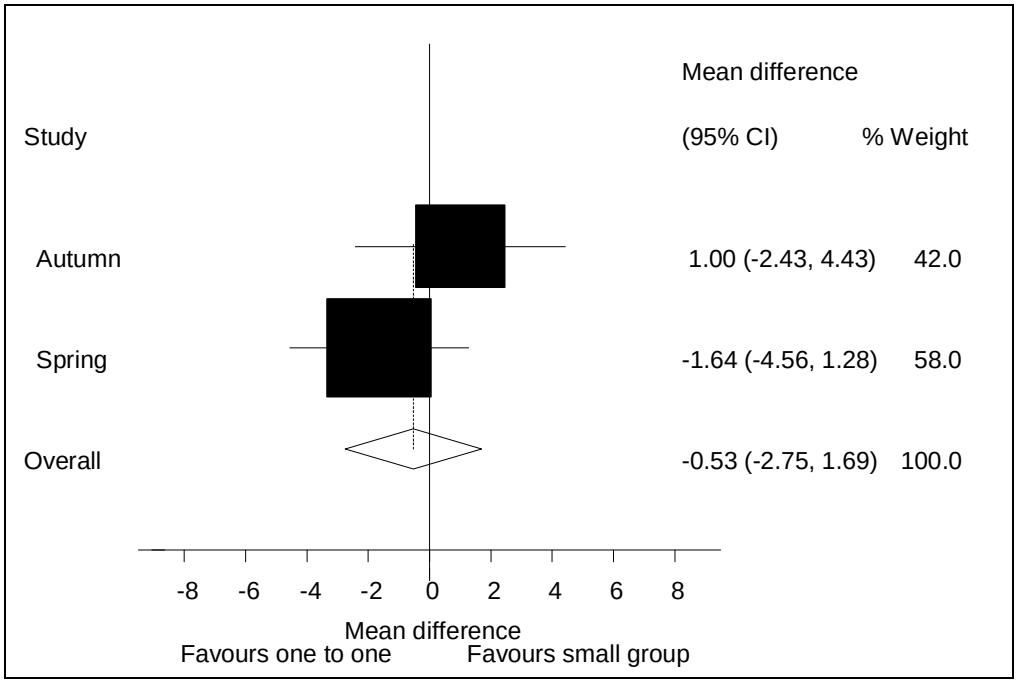
April testing

The mean PIM 6 score for the children receiving NC individually was 22.8 (SD 3.4) and for the children receiving adapted NC in triplets was 19.2 (SD 5.4). The results highlight the finding that children who received adapted NC in triplets scored slightly lower, although not statistically significantly lower, on the PIM 6 mathematics test compared with children who had received NC intervention individually (-1.6 95% CI -4.6 to 1.3, p=0.23). The effect size was -0.32 (95% CI -1.18 to 0.54) (Table 4.6).

Overall

In figure 4.2 we combine the two analyses of one-to-one teaching versus teaching in triplets. As the figure demonstrates, the pooling of the two trials shows no statistically significant difference between the scores of the children taught individually or in triplets in terms of PIM 6 scores.

Figure 4-2: Forest plot of NC one-to-one vs. one-to-three teaching



In Figure 4-2 we show graphically the results of the autumn and spring results. There are two study plots: the first shows the estimated effect of the study with horizontal lines showing how uncertain we are about the results. There is uncertainty about the results because of the relatively small sample size and so we can see that the horizontal lines cross over the vertical axis which is zero. In other words no statistically significant effect was observed. The second plot is on the other side of no effect but its horizontal lines also pass through zero or the no effect vertical line. Finally the two studies are combined in the diamond figure

at the bottom of the graph. The peak of the diamond is just past the left hand side of the line of zero indicating a small benefit, in favour of one-to-one which is not statistically significant because part of the diamond passes over the vertical line of no effect.

Secondary outcome

The secondary outcome was score on Sandwell B test. This measure was not undertaken blind or marked blind to group allocation.

Table 4-7: Trial 2: Triplets Secondary outcomes

Outcome	Singles	Triplets	Estimate*	ES
Sandwell B (January), mean (SD)**	50.6 (8.5) [n=9]	50.6 (11.5) [n=44]	-0.6 (-9.5 to 8.4) [n=53]	-0.05 (-0.77 to 0.66)
Sandwell A (April), mean (SD)	54.6 (12.7) [n=7]	49.3 (12.7) [n=44]	-1.2 (-8.6 to 6.2) [n=51]	-0.10 (-0.90 to 0.70)

* Analyses were adjusted for baseline Sandwell A test scores, age, gender, free school meals and the clustering within schools

**Analysis excludes children who could not be randomized to autumn term.

The effect size for the January testing was -0.05 (-0.77 to 0.66) and for the April testing was -0.10 (-0.90 to 0.70). However, we are unable to rule out ascertainment bias as a potential threat to the reliability of this measure.

4.4: Discussion

This trial is very small; consequently any conclusions based on the data must be treated very cautiously. There were no statistically significant differences in impact between the groups. To resolve the issue of which is most effective: individual, pair or triplet delivery, we would recommend a large rigorous trial comparing NC delivered one-to-one with adapted NC being delivered in pairs and triplets.

The trial has some strengths notably robust randomization and blinded follow-up; however, the key weakness – small sample size – precludes us from making any strong conclusions about the relative efficacy of triplet teaching.

4.5: Conclusions

This small study demonstrated that there is no evidence of any difference in impact between children taught individually or in triplets. We are cautious about generalising this finding because of the small sample size.

Chapter 5: Impact of Every Child Counts: Secondary analyses

Key summary points

- The key findings are based on four quasi-experiments using two designs – interrupted time series (one QED) and case control design (three QEDs). Quasi-experimental designs are inherently less rigorous than randomized controlled trial designs. We designed and conducted the QEDs to be as rigorous as possible. However, the findings should be interpreted with caution.
- The secondary analyses investigated short- and medium-term impact of Every Child Counts (ECC) compared with no ECC at the school level.
- The interrupted time series analysis showed a small statistically significant positive effect. However there was a lack of a clear linear trend from which to estimate the discontinuity. Furthermore, the results for English were similar to those for mathematics suggesting some alternative explanation other than NC.
- The case control studies tended to show a small positive effect, usually not statistically significant, on mathematics scores and a small negative effect (not statistically significant) on English scores.
- In summary, the findings from the secondary analyses strand of the evaluation are inconclusive as we could not detect an impact of ECC on KS1 results.

5.1: Introduction

In order to provide impact data to assess the effectiveness of the Every Child Counts (ECC) programme on improving children's attainment in mathematics at school level we carried out a series of secondary data analyses. The National Pupil Database (NPD) was used to carry out four quasi-experiments using two quasi-experimental designs: interrupted time series (ITS) design and case control design (CC).

The secondary analyses used data from all of the intervention children in the 2008-9, 2009-10 cohort schools (with the exception of the schools taking part in Trial 2), as well as data from all children in these schools not exposed to the ECC intervention, historical data from the same schools and data from matched comparison schools derived from the National Pupil Database. We assessed the impact of ECC compared with non treated controls using 2009 KS1 outcomes.

We received approval for our Protocol from the DfE and the Steering Group. The secondary analyses protocol includes dates approval was received from the University of York HSSEC, the DfE and the Steering Group (Torgerson et al, 2011c).

5.2: Design and methods

5.2.1: Interrupted time series (ITS) design

We used historical data to set up an interrupted time series (ITS) approach with each school acting as its own control. In an ITS design a group of participating schools is tested repeatedly both before and after the introduction of an intervention, in this case before and after the introduction of ECC. In essence this is a single-group, pre- and post-test design with multiple before and after measurements which enables control of some of the bias that may occur due to regression to the mean or temporal effects.

We plotted whole school KS1 results in mathematics and English for the three data points preceding the implementation of ECC in the 2008-9 cohort schools (2006, 2007 and 2008). KS1 English points score was used in order to provide evidence that we were measuring a real effect. We ran equivalent models using English as an outcome. We would expect the estimated effect of the mathematics intervention to be lower for English than for mathematics (although we hypothesised that some effect may be present for English as an outcome, as the intervention involved work on vocabulary.) This then provided a baseline for data points *after* implementation of ECC in 2008-9 in 2009 and 2010.

5.2.2: Multilevel models

We ran multilevel models in order to estimate the effect of the intervention. A school's average KS1 score in a given year was the first level; schools were the second level. As previously mentioned, we controlled for any trends over time by including academic year in the models. To test whether the relationship was approximately linear or should be modelled as a curve, ordinary least squares (OLS) regression was used modelling it as a linear and a quadratic relationship. For both mathematics and English the adjusted R-squared decreased with the inclusion of the quadratic term, suggesting the linear model was preferable. A dummy variable for pre- and post-intervention was included in the models, with the coefficient providing estimates of the effect of the intervention.

The following multilevel models were constructed (for both mathematics and English KS1 outcomes) in order to give us a number of estimates of the effect of ECC (as an illustration the equation for model 3 is included):

1. A dummy for pre- and post-intervention, a fixed linear relationship with academic year, and a random effect at the first level

2. As (1) above but introducing a random effect at the school level, essentially allowing schools to have different mean KS1 point scores but assuming the effect of adopting ECC was the same between schools

3. As (2) above but allowing the coefficient of academic year to vary at school level, essentially allowing the underlying trend of improving or worsening KS results over time to differ between schools (see Figure 5-1 for the equation of this model)

Figure 5-1: The multilevel model equation for model 3

$$\text{KS1\_Maths\_Points}_{ij} \sim \text{N}(XB,\ \Omega)$$

$$\text{KS1\_Maths\_Points}_{ij} = \beta_{0ij}\text{cons} + \beta_1\text{postIntervention}_{ij} + \beta_{2j}\text{acYear}_{ij}$$

$$\beta_{0ij} = \beta_0 + u_{0j} + e_{0ij}$$

$$\beta_{2j} = \beta_2 + u_{2j}$$

$$\begin{bmatrix} u_{0j} \\ u_{2j} \end{bmatrix} \sim \text{N}(0,\ \Omega_u) \ :\ \Omega_u = \begin{bmatrix} \sigma^2_{u0} & \\ \sigma_{u02} & \sigma^2_{u2} \end{bmatrix}$$

$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim \text{N}(0,\ \Omega_e) \ :\ \Omega_e = \begin{bmatrix} \sigma^2_{e0} \end{bmatrix}$$

In this way we set up a quasi-experiment in which it was possible to see whether ECC has had an impact on whole school attainment in the Yr 1 ECC cohort. It is important to note that this analysis takes into account the fact that not all Numbers Count Teachers had received accreditation during the period 2008-9. With two post-intervention time points we were able to make some estimate of a whole-school treatment effect in KS1 outcomes.

If the plot of the dependent variable (in this case KS1 outcomes in mathematics and English) shows a change in level or direction at the point of intervention (either immediately after or delayed), and potentially confounding variables have been minimised due to multiple observations (in this case use of multiple schools), then it is possible to ascribe a causal relationship between the intervention and the dependent variable (in this case KS1 outcomes). With two post-intervention time points we can make some estimate of a whole-school treatment effect in KS1 outcomes. However, it should also be noted that the ITS design does not permit such a strong causal relationship to be established as the more rigorous randomized controlled trial design. For example, other changes in policy occurring at the same time as ECC could have confounded the results of the evaluation using this method. In education, this is a real threat due to multiple policy changes. This is not the case using a randomized controlled trial design. Therefore, the results from the ITS analyses in the secondary analyses should be treated with caution.

Our sample for the interrupted time series (ITS) design was all schools that introduced the ECC intervention in 2008-9 and continued it in 2009-10. We removed schools that took part in Trial 2 of the ECC evaluation as they used a different form of the intervention (teaching in pairs and triplets, rather than one-to-one). To model the impact of ECC the data were aggregated to school level for each year. Data from years 2006, 2007 and 2008 provided baseline data prior to the introduction of the intervention in 2008 and post-intervention data were used from years 2009 and 2010. We needed to control for any trends over time (for example if KS1 results in general improve over time then any increase in results may be confused with the effect of ECC). We controlled for this by including the academic year in the models.

5.2.3: Case control design

We also carried out three case control design (CCD) quasi-experiments. In a CCD participants are identified as receiving a specific intervention and compared with a control group of participants without the intervention. In this instance the KS1 outcomes for schools already implementing ECC were compared with the KS1 outcomes for matched control schools. As with the ITS design, the case control design provides a mechanism for establishing a causal link between ECC and KS1 outcomes, but, due to the limitations associated with matching, the causal link is not as strong as that provided by the more rigorous randomized controlled trial design. Because controls have not been randomly allocated there is a possibility that selection bias will affect the results. This is because control schools may have subtle but important differences that could affect outcomes.

In the first case control design we compared KS1 outcomes for schools in the 2008-9 cohort with comparison schools matched using the nationally available National Pupil Database (NPD) data. The matched comparison schools started to implement the intervention in the period 2009-10. We matched the schools based on Foundation Stage Profile (FSP) data and other available variables such as proportion of children eligible for free school meals. We used the KS1 scores in 2009 and 2010 to assess the differences in outcome between the two groups of schools – ECC 2008 cohort and matched comparison schools. Comparison of ECC children's KS1 outcomes with the outcomes of the children in the matched comparison group, including subgroup variation was possible. Subgroup analyses included lowest band of attainers at FSP and FSM status. Both descriptive and inferential statistics are presented. *Multiple regression* and *multilevel modelling* were used to investigate the effect of ECC, whilst controlling for any measured differences between the two groups. In order to control for researcher bias the researchers were blind to intervention and comparison groups until after the interpretation of the results.

Post-hoc analysis

Subgroup analyses were performed in order to see if the intervention effect varied for different subgroups. We compared low, medium and high groups of attainers at FSP and low, medium and high subgroups of IDACI scores. In both comparisons the data were split into equal thirds.

In the second CCD we repeated the analyses of the first CCD, but this time we compared 2009-10 KS1 outcomes for schools. This enabled us to investigate the effect (school level) of schools having taken part in ECC for two years over and above the effect of taking part for 1 year.

In the third case control study pre- and post-test data for ECC 2009-10 cohort were compared with pre- and post-test data for a matched comparison group. We compared KS1 outcomes for schools in the 2009-10 cohort with control schools matched using the nationally available National Pupil Database (NPD) data. The comparison schools were matched on Foundation Stage Profile (FSP) data and other available variables such as proportion of children eligible for free school meals. We used the KS1 scores in 2010 to assess the differences in outcome between the two groups of schools. We compared ECC children's KS1 outcomes with the outcomes of the children in the matched comparison group. Subgroup analyses include the lowest band of attainers at FSP and FSM status.

For all secondary analyses (ITS and CCD 1, 2 and 3) the following procedures were undertaken prior to data analysis.

Key stage 1 mathematics

The KS1 mathematics levels were converted to points using the official[4] guidance:

W = Working towards level 1 = 3 points

1 = Achieved Level 1 = 9 points

2c = Achieved Level 2c = 13 points

2b = Achieved Level 2b = 15 points

2a = Achieved Level 2a = 17 points

3 = Achieved Level 3 = 21 points

4+ = Achieved Level 4 or above = 27 points

5.3: Results

5.3.1: Interrupted time series design

Descriptive statistics

175 schools were included in the sample (schools that introduced ECC in 2008-9 and continued with it in 2009-10).

Table 5-1 presents descriptive statistics for each of the 5 years of data based on pupil level data. The sample includes around 7000 pupils in each year. There was an overall tendency for the KS1 mathematics and English scores to increase over time. This is demonstrated more clearly in Figure 5-2. In order to control for this tendency for KS1 results to increase over time we included year as a variable in the model. However, the trend is not clearly linear and the interpretation of a post-intervention difference as a causal effect in this case is problematic.

There was a year on year increase or decrease for a number of variables. SEN involvement increased over time, as did the IDACI scores (higher IDACI scores indicate higher deprivation). The proportion of pupils with English as a first language decreased over time. The proportion of white students decreased over time, whilst the proportion of Asian students increased over time.

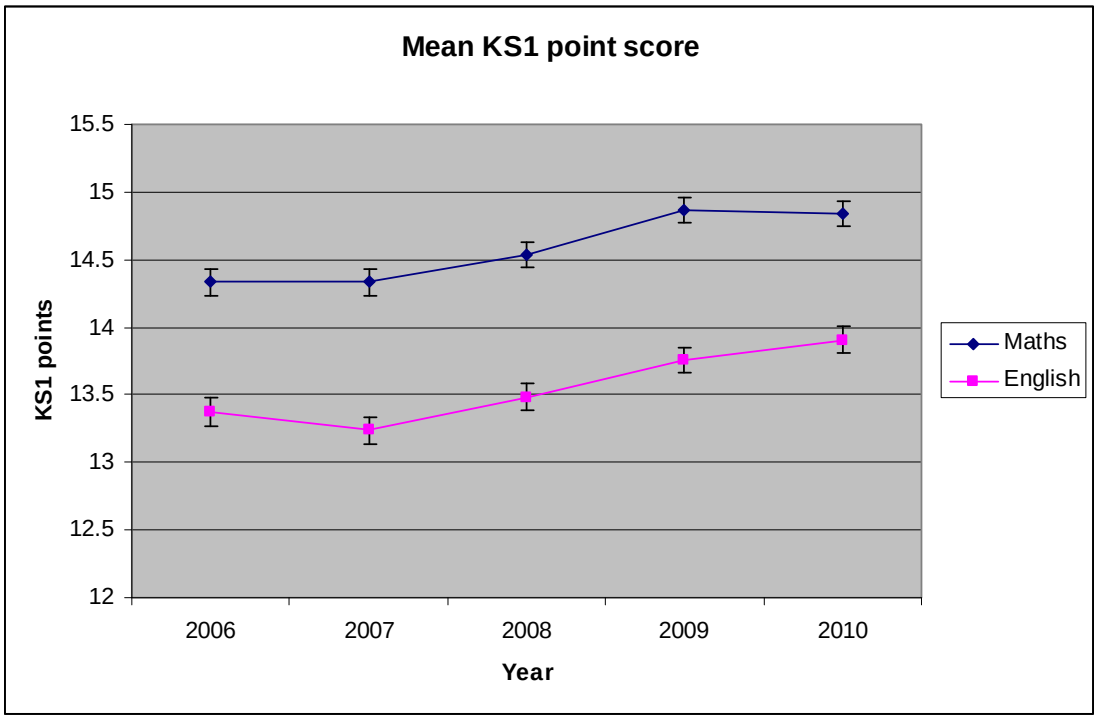
[4] Measuring Progress at Pupil, School and National levels DFE (2009). www.education.gov.uk/research/data/uploadfiles/DCSF-RTP-09-02.pdf

Table 5-1: Descriptive statistics

	Pre-intervention			Post-intervention	
	2005-06	2006-07	2007-08	2008-09	2009-10
Number of pupils	7020	7184	7047	7095	7215
Mean KS1 Maths points (SD)	14.333 (4.040)	14.333 (4.059)	14.540 (3.928)	14.862 (3.794)	14.841 (3.746)
Mean KS1 English points (SD)	13.368 (4.426)	13.235 (4.341)	13.485 (4.222)	13.757 (4.157)	13.905 (4.150)
Mean FSP Maths score (SD)	19.545 (5.552)	18.531 (5.598)	18.028 (5.116)	18.230 (5.174)	18.280 (4.969)
Mean IDACI (SD)	0.387 (0.199)	0.387 (0.195)	0.406 (0.199)	0.407 (0.199)	0.412 (0.199)
Mean Age* (SD)	5.486 (3.516)	5.526 (3.476)	5.523 (3.479)	5.554 (3.479)	5.467 (3.468)
% boys	51.8%	51.9%	50.3%	51.2%	51.9%
% SEN involvement	25.6%	28.9%	29.5%	29.6%	30.3%
% English first language	72.5%	70.4%	70.0%	67.7%	66.5%
% White	65.2%	64.1%	63.5%	61.8%	59.7%
% Asian	19.3%	20.1%	20.4%	21.2%	21.8%
% Black	9.7%	10.1%	9.9%	10.6%	11.3%
% Chinese	0.4%	0.3%	0.3%	0.3%	0.4%

* number of months over 6 years at the end of academic year

Figure 5-2: Mean KS point scores, split by year (based on pupil data)



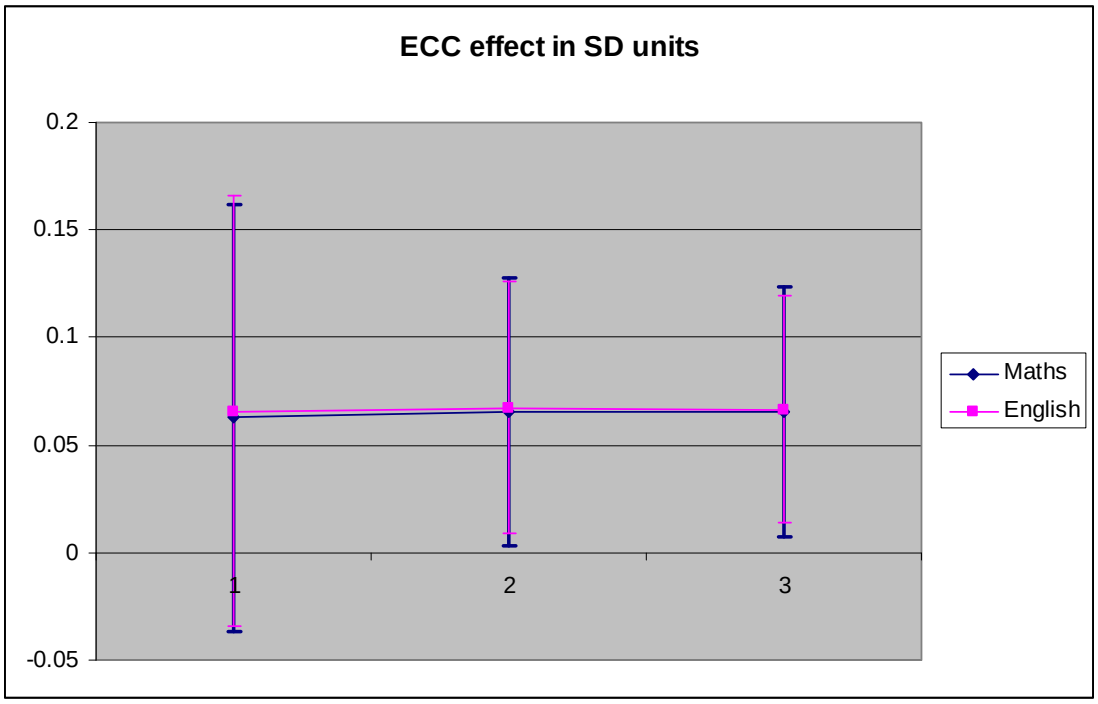
Results of multilevel modelling

The results from running the 3 models are presented in Table 5-2 and the effects (along with the 95% confidence intervals) are presented in Figure 5-3. Two of the three models showed a small statistically significant discontinuity in the trend of around 0.25 KS1 points (with mathematics as an outcome) which translates to an effect size of 0.066 in standard deviation units. However using English as an outcome the effects are similar (this is particularly clear in Figure 5-3). Further post-hoc analysis introducing various relevant covariates into the models continued to show the mathematics and English effects as being equal. Given the lack of a clear linear trend from which to estimate the discontinuity, any causal effect should be inferred with caution.

Table 5-2: Results of multilevel models showing estimates of the ECC effect

	ECC effect in KS points (SE)		ECC effect in SD units (SE)	
Model	Maths	English	Maths	English
1	0.242 (0.191)	0.273 (0.207)	0.063 (0.050)	0.066 (0.050)
2	0.253 (0.120)	0.280 (0.121)	0.066 (0.031)	0.068 (0.029)
3	0.252 (0.112)	0.276 (0.109)	0.065 (0.029)	0.067 (0.026)

Figure 5-3: Results of multilevel models showing estimates of the ECC effect



5.3.2: Case control design 1

In Figure 5-4 and Figure 5-5 below we present distributions of prior academic achievement (as measured by Foundation Stage Profile (FSP) mathematics score) and deprivation (as measured by IDACI) for the intervention and comparison group. As the distributions for the two groups were similar we decided that propensity matching (to prevent our analyses from being compromised by non-overlapping comparison groups) was not required.

Figure 5-4: Distribution of prior mathematics ability (FSP scores), split by 2008 cohort and matched comparison group

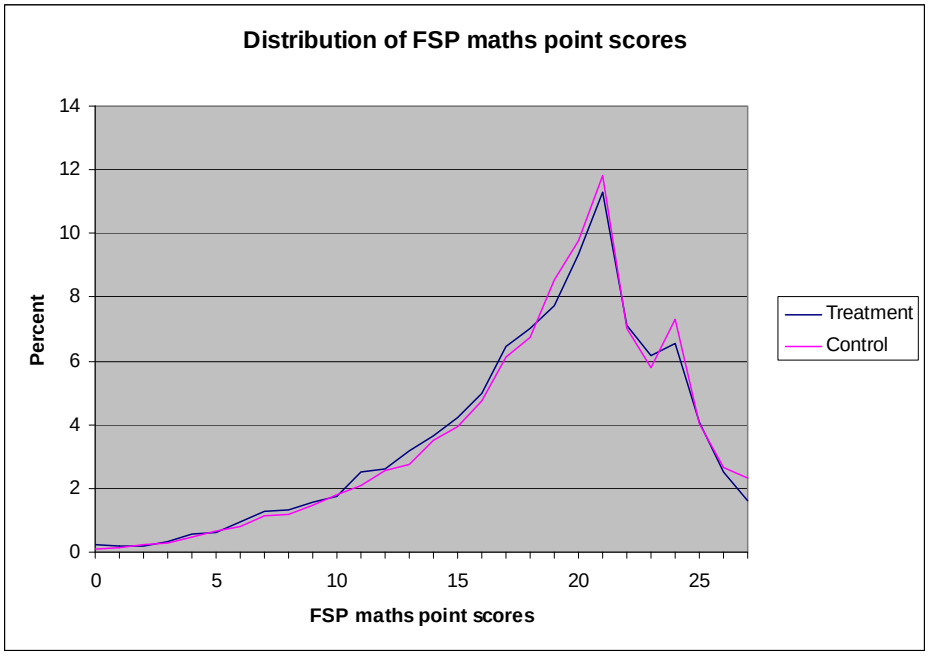
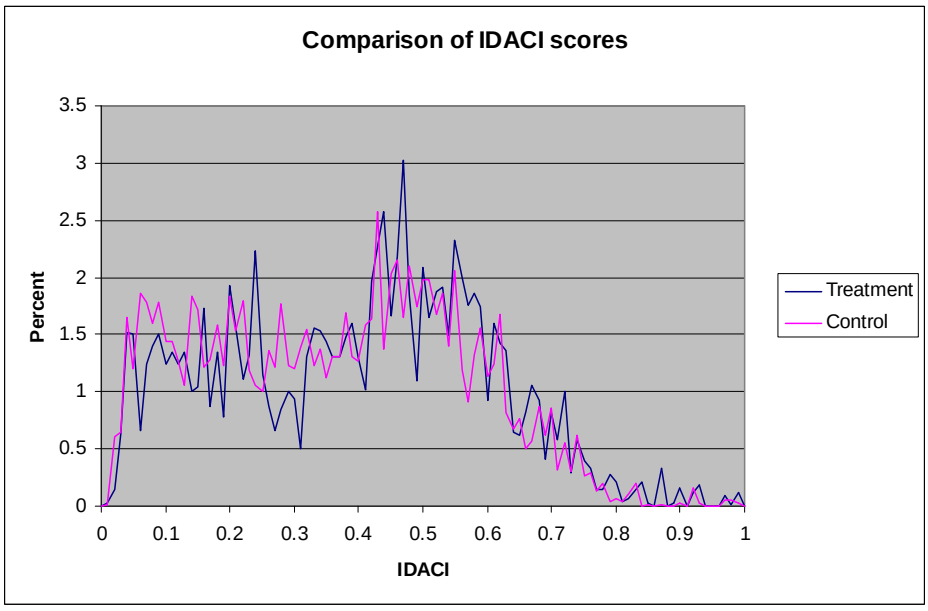


Figure 5-5: Distribution of deprivation scores (IDACI scores), split by 2008 cohort and matched comparison group



Descriptive statistics are presented comparing the 2008 cohort with the matched comparison group. We present the results of the following models in order to investigate the effect of introducing various covariates including: Null, FSP plus all pupil level and school level variables.

Descriptive statistics

Table 5-3 presents descriptive statistics comparing intervention and comparison groups. The comparison group had, on average, slightly fewer students from deprived backgrounds and slightly younger students and the intervention group, on average, had pupils with slightly lower FSP scores. However these differences were small. In terms of the proportion of boys, SEN involvement and ethnic diversity, the groups were similar, the biggest difference being that the intervention group had around 2.5% more students with SEN involvement.

Table 5-3: Descriptive statistics

	Intervention	Comparison
Number of pupils	9319	16921
Mean FSP Maths score (SD)	18.305 (5.181)	18.603 (5.087)
Mean IDACI (SD)	0.396 (0.203)	0.365 (0.199)
Mean Age* (SD)	5.506 (3.473)	5.483 (3.472)
% boys	51.8%	50.7%
% SEN involvement	29.8%	27.2%
% English first language	68.8%	69.7%
% White	63.0%	62.5%
% Asian	20.9%	22.2%
% Black	9.9%	8.7%
% Chinese	0.3%	0.3%

* number of months over 6 years at end of academic year

In summary these results show that the two groups were broadly similar but minor differences existed and these differences were controlled for in the analyses.

Results of multiple regression and multilevel modelling

Table 5-4 and Table 5-5 present the results of all our models that estimated the effect of the ECC intervention on the schools involved. Table 5-4 presents estimates using KS1 mathematics points as the outcome and Table 5-5 presents results from equivalent models using KS1 English points as an outcome. Results from both multiple regression and multilevel modelling are presented.

Table 5-4 shows the effect of ECC using KS1 mathematics points as an outcome. The Null model (1) shows an ECC effect of 0.06 in key stage mathematics points, i.e., there was very little difference between the KS1 mathematics results of the ECC 2008 cohort and matched comparison group. With an R-squared of zero this is a poorly fitting model. However the adjusted R-squared results for models 2 to 4 explain a higher proportion of variance in the outcomes (almost 50%) and are therefore better models. The six estimates of the ECC effect using these three models with the two methods of analyses (multiple regression and multi-level modelling) range from 0.196 to 0.275 KS points. These results give similar estimates of the ECC effect of a little over 0.2 KS points.

The effects are also presented in SD units and show a range of effects that are broadly similar, around 0.06 SD units. These effects are relatively small but they are whole-school effects. Figure 5-6 shows the ECC effects in SD units graphically – again it is clear that models 2-4 estimate the ECC effect as being around 0.06 SD units.

Table 5-4: Estimates of the effect of ECC using KS1 mathematics points as the outcome

	Multiple Regression results				Multi-level Model results	
Model	ECC effect in KS points (SE)	ECC effect in SD units (SE)	Adjusted R-squared	Number of pupils	ECC effect in KS points (SE)	ECC effect in SD units (SE)
1	0.060 (0.050)	0.016 (0.013)	0.000	26207	0.088 (0.106)	0.023 (0.027)
2	0.196 (0.038)	0.051 (0.010)	0.428	24826	0.233 (0.105)	0.060 (0.027)
3	0.252 (0.037)	0.065 (0.010)	0.476	24504	0.275 (0.099)	0.071 (0.026)
4	0.219 (0.036)	0.057 (0.009)	0.494	24504	0.238 (0.088)	0.062 (0.023)

Table 5-5: Estimates of the effect of ECC using KS1 English points as the outcome

	Multiple Regression results				Multi-level Model results	
Model	ECC effect in KS points (SE)	ECC effect in SD units (SE)	Adjusted R-squared	Number of pupils	ECC effect in KS points (SE)	ECC effect in SD units (SE)
1	-0.235 (0.054)	-0.056 (0.013)	0.001	26203	-0.212 (0.116)	-0.051 (0.028)
2	-0.088 (0.040)	-0.021 (0.010)	0.441	24822	-0.049 (0.114)	-0.012 (0.027)
3	0.008 (0.038)	0.002 (0.009)	0.523	24501	0.021 (0.105)	0.005 (0.025)
4	-0.025 (0.037)	-0.006 (0.009)	0.538	24501	-0.011 (0.094)	-0.003 (0.023)

Figure 5-6: Estimates of the effect of ECC using KS1 mathematics points as the outcome

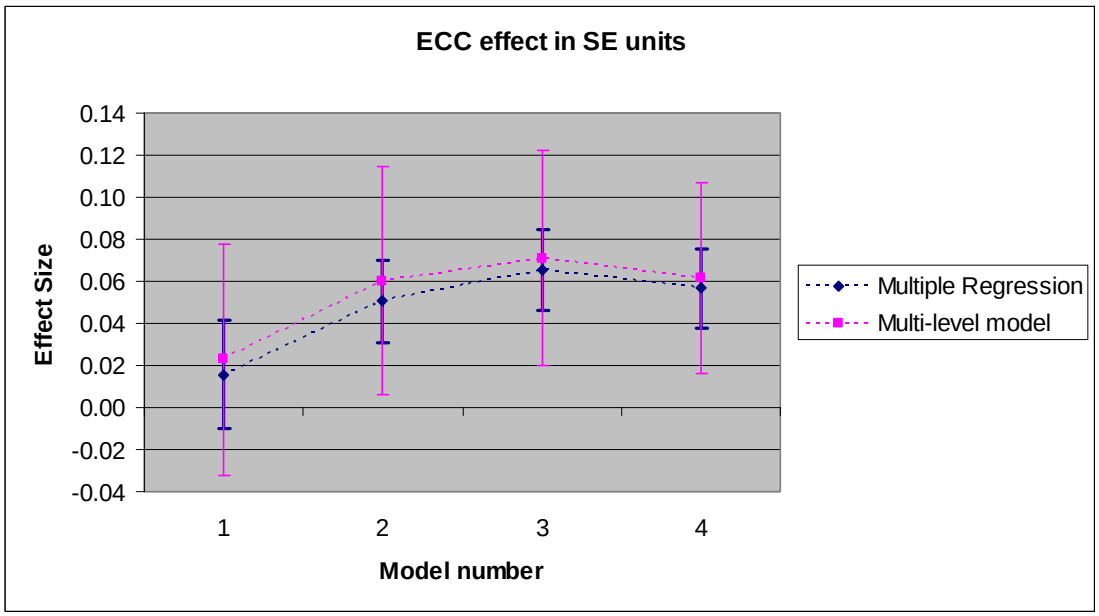


Table 5-5 shows the estimates of the ECC effect with English KS1 as the outcome. Again models 2-4 are the better fitting models and they show effect sizes that are greatly reduced to almost zero, in comparison with the results from Table 5-4, which were for mathematics as an outcome. We would expect the ECC effect on English to be smaller than the effect on

mathematics as it is a numeracy intervention. These results give us more confidence that the effects measured in Table 5-4 are true effects.

Subgroup analyses

Subgroup analyses were performed to investigate whether the ECC effect varied for different groups. Subgroups were formed based on prior mathematics attainments (FSP mathematics scores) and deprivation (IDACI scores). Ranges of scores were chosen in order to split the pupils into three groups (with lower, intermediate and higher scores) of equal size for both FSP and IDACI. The analyses were then run using just Model 4 (using all covariates) and multiple regression for simplicity. The results are presented in Table 5-6 and Table 5-7 below and presented graphically in Figure 5-7 and Figure 5-8.

Table 5-6 shows the results of the subgroup analyses based on FSP. Group 1 contains pupils with the lowest FSP scores, group 2 the intermediate scores and group 3 pupils with the highest FSP scores. We can see that the ECC effect is largest in the group with the lowest FSP scores and the effect decreases with the intermediate group and reduces to almost zero with the group with the highest FSP scores. The reduction in effect is presented in Figure 5-7 along with the 95% confidence intervals. Figure 5-7 shows that if we take the confidence intervals into account the ECC effect seems to be bigger for the subgroup with the lowest FSP scores than for the subgroup with the highest FSP scores (based on FSP mathematics scores). The differences were statistically significant.

Table 5-6: FSP

		Range			KS1 Maths points		ECC effect			
Group	n	Min	Max	Mean	Mean	SD	ECC effect in KS points	Standard Error	ECC effect in SD units	Standard Error
1	8665	0	17	12.835	12.204	3.725	0.381	0.068	0.102	0.018
2	9024	18	21	19.717	15.452	2.721	0.186	0.056	0.068	0.021
3	7137	22	27	23.841	17.576	2.737	0.074	0.064	0.027	0.023

Figure 5-7: FSP

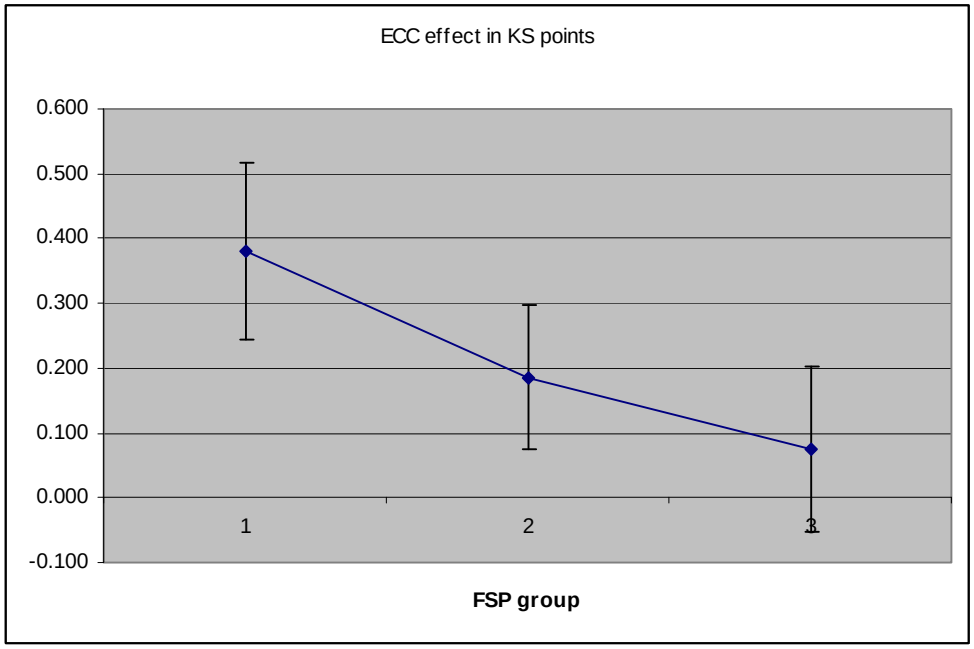
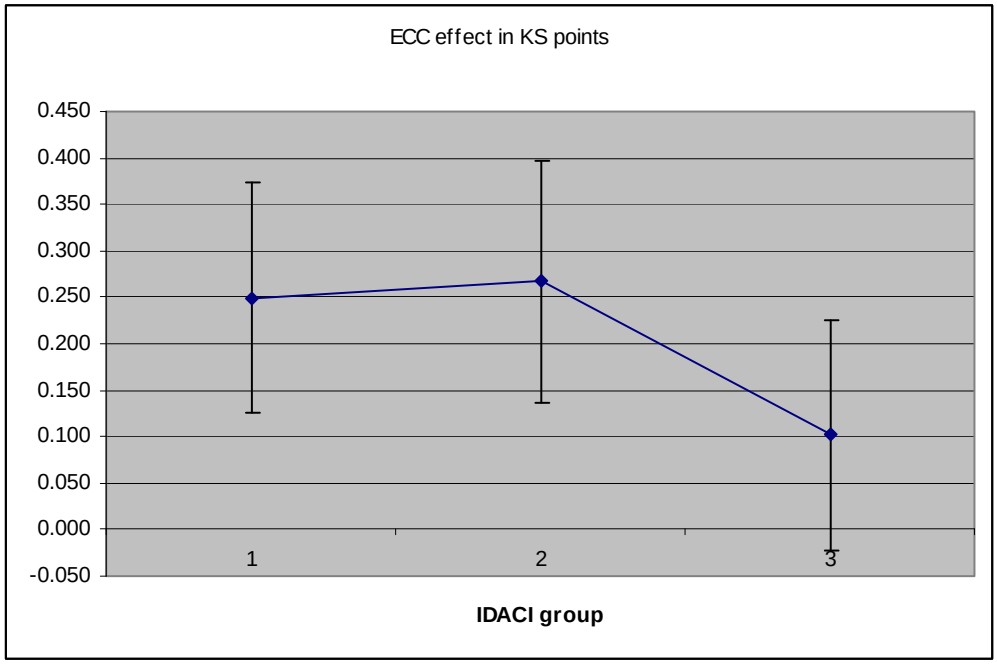


Table 5-7 shows the results of the subgroup analyses based on IDACI scores. Group 1 contains pupils with the lowest IDACI scores (the least deprived), group 2 the intermediate scores and group 3 pupils with the highest IDACI scores (i.e. the most deprived students). The ECC effect may reduce as the measure of deprivation increases. This means that ECC may have a smaller impact on children with the highest deprivation scores. However as there are no statistically significant differences between the groups this could be due to chance.

Table 5-7: IDACI

		Range			KS1 Maths points		ECC effect			
Group	n	Min	Max	Mean	Mean	SD	ECC effect in KS points	Standard Error	ECC effect in SD units	Standard Error
1	8652	0.011	0.264	0.143	15.569	3.779	0.249	0.062	0.066	0.016
2	8654	0.264	0.481	0.386	14.560	3.870	0.267	0.065	0.069	0.017
3	8645	0.481	0.990	0.599	14.389	3.788	0.102	0.062	0.027	0.016

Figure 5-8: IDACI



5.3.3: Case control design 2

CCD2 involved the same group of schools as CCD1; however, we used 2009/10 outcome data rather than data for 2008/09. In design 2 we investigated the school level effect of schools having taken part in ECC for two years over and above the effect of taking part for one year. We therefore compared the 2009/10 outcomes between the two groups.

The same analyses used in CCD1 were repeated in CCD2. Descriptive statistics are presented comparing the two year intervention group and the one year intervention group. *Multiple regression* and *multilevel modelling* were used to investigate the effect of running the intervention for two years compared to one year, whilst controlling for any measured differences between the two groups. We repeated the same models used in CCD1 and used the same outcome measures and covariates.

Descriptive statistics

Table 5-8: Descriptive statistics

	Intervention 2 years	Intervention 1 year
Number of pupils	9489	17910
Mean FSP Maths score (SD)	18.437 (4.918)	18.470 (4.934)
Mean IDACI (SD)	0.400 (0.204)	0.368 (0.199)
Mean Age* (SD)	5.469 (3.456)	5.454 (3.475)
% boys	51.8%	51.1%
% SEN involvement	29.4%	27.4%
% English first language	68.2%	69.3%
% White	60.7%	62.5%
% Asian	21.2%	21.9%
% Black	10.7%	8.9%
% Chinese	0.4%	0.3%

* number of months over 6 years at the end of academic year

Table 5-8 presents descriptive statistics comparing schools that took part in ECC for two years and schools that were involved for one year. There were nearly twice as many pupils in the one year group compared with the two year group. The two groups had very similar characteristics, with most means or proportions matching very well. The year two group had slightly more deprived students, slightly more SEN involvement and a slightly wider ethnic mix, but the differences were small. These results show that the two groups were broadly similar but minor differences existed and these differences were controlled for in the analyses.

Results of multiple regression and multilevel modelling

Table 5-9 and Table 5-10 present the results of all our models that attempted to estimate the effect of being in ECC for two years over and above the effect of being in for one year. Table 5-9 presents estimates using KS1 mathematics points as the outcome and Table 5-10 presents results from equivalent models using KS1 English points as an outcome. Results from both multiple regression and multilevel modelling are presented.

Table 5-9: Estimates of the effect of 2 years of ECC over and above 1 year of ECC using KS1 mathematics points as the outcome

	Multiple Regression results				Multi-level Model results	
Model	Effect of second year of ECC (SE)	Effect in SD units (SE)	Adjusted R-squared	Number of pupils	Effect of second year of ECC (SE)	Effect in SD units (SE)
1	0.015 (0.048)	0.004 (0.013)	0.000	27338	0.003 (0.108)	0.001 (0.029)
2	-0.019 (0.037)	-0.005 (0.010)	0.411	26000	-0.009 (0.101)	-0.002 (0.027)
3	0.046 (0.035)	0.012 (0.009)	0.460	25606	0.038 (0.094)	0.010 (0.025)
4	0.073 (0.035)	0.019 (0.009)	0.475	25606	0.065 (0.085)	0.017 (0.023)

Table 5-10: Estimates of the effect of 2 years of ECC over and above 1 year of ECC using KS1 English points as the outcome

	Multiple Regression results				Multi-level Model results	
Model	Effect of second year of ECC (SE)	Effect in SD units (SE)	Adjusted R-squared	Number of pupils	Effect of second year of ECC (SE)	Effect in SD units (SE)
1	-0.087 (0.053)	-0.021 (0.013)	0.000	27341	-0.110 (0.121)	-0.027 (0.029)
2	-0.141 (0.040)	-0.034 (0.010)	0.428	25993	-0.132 (0.114)	-0.032 (0.027)
3	-0.049 (0.037)	-0.012 (0.009)	0.510	25599	-0.068 (0.105)	-0.016 (0.025)
4	-0.017 (0.037)	-0.004 (0.009)	0.524	25599	-0.030 (0.097)	-0.007 (0.023)

Figure 5-9: Estimates of the effect of 2 years of ECC over and above 1 year of ECC using KS1 mathematics points as the outcome

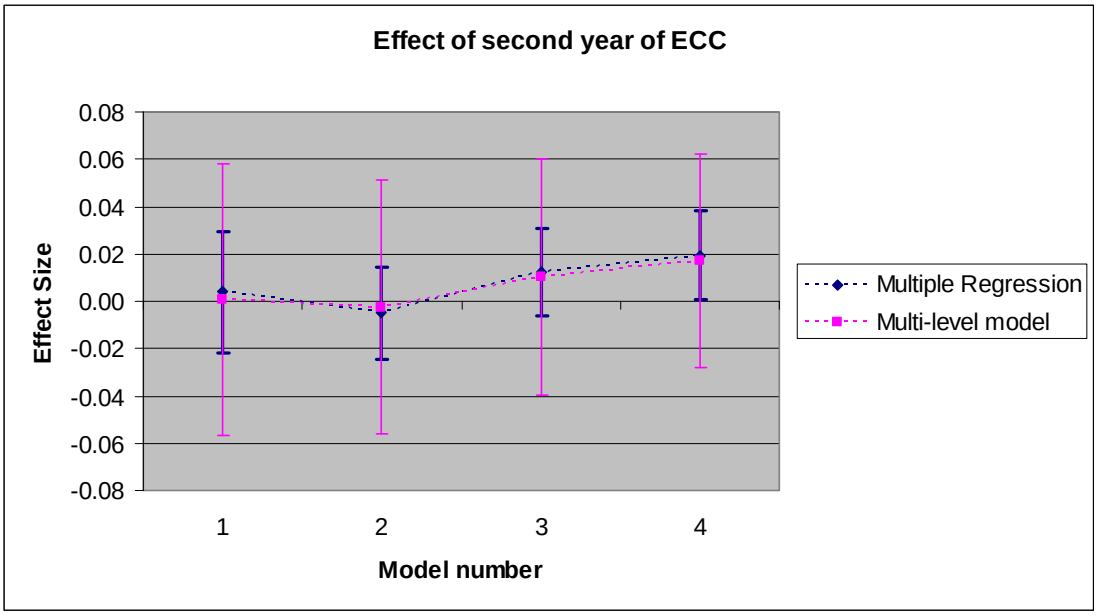


Table 5-9 shows the effect using KS1 mathematics points as an outcome. The null model (1) shows an effect of 0.015 in KS1 mathematics points, demonstrating there was very little difference between the KS1 mathematics results of the intervention group and the control group. With an R-squared of zero this is a poorly fitting model. However the adjusted R-squared results for models 2 to 4 explain a higher proportion of variance in the outcomes (almost 50%) and are therefore better models. The six estimates of the effect using these three models with the two methods of analyses (multiple regression and multi-level modelling) range from -0.019 to 0.073 KS points which were very small effects and for all but one result they were not significantly different from zero. The effects in SD units range from -0.005 to +0.019 as illustrated in Figure 5-9.

Table 5-10 shows the estimates of the ECC effect with English KS 1 as the outcome. Again there were no effect sizes that were significantly different from zero with the exception of Model 2.

Subgroup analyses

As we did not detect a main effect it was inappropriate to undertake any sub-group analyses.

5.3.4: Case control design 3

The intervention group in CCD3 consisted of the schools that took part in ECC in 2009-10 (with the exception of schools involved with Trial 2 as they received the pair and triplets intervention). In order to estimate the ECC effect, the KS 1 results from the intervention schools were compared to the results of matched control schools that had not taken part in ECC (schools that took part in ECC in 2008-09 but did not continue with ECC in 2009-10 were removed from the sample before matching took place).

The comparison group was identified using propensity matching. We used the dummy variable (1=intervention group, 0=control group) as the dependent variable and covariates aggregated to school level as independent variables and ran a logistic regression to calculate a propensity score for each school. This gave the probability that a school with those scores is a member of the intervention group. For each school in the intervention group we considered the two schools with the next highest and next lowest propensity scores. If neither of these schools was part of the intervention group, the school with the nearer propensity score was added to the control group as a match. If only one of these schools was part of the intervention group then the other school was added to the comparison group as a match. In the event that both schools were also part of the intervention group, no match was returned and the school was omitted from the analysis. Using this method it was possible for the same non-ECC school to be returned as the matching comparison school for two different intervention schools – this was allowed and no weighting was applied.

Descriptive statistics

The propensity matching process matched 573 out of the total 578 ECC schools to 559 non-ECC schools. Descriptive statistics comparing the two groups at the pupil level are presented in Table 5-11: Descriptive statistics below. The two groups were broadly similar, with a few small differences. The intervention group had, on average, slightly lower prior attainment scores and came from slightly more deprived backgrounds. The intervention group also had a higher proportion of pupils with SEN involvement and a higher proportion of Asian pupils. We controlled for these small differences in the models.

Table 5-11: Descriptive statistics

	ECC group	Control group
Number of pupils	24908	24710
Mean FSP Maths score (SD)	18.415 (4.955)	19.012 (4.921)
Mean IDACI (SD)	0.379 (0.200)	0.368 (0.207)
Mean Age* (SD)	5.455 (3.472)	5.457 (3.475)
% boys	51.4%	50.5%
% SEN involvement	28.3%	26.8%
% English first language	68.8%	69.9%
% White	62.0%	62.2%
% Asian	21.6%	19.4%
% Black	9.5%	10.3%
% Chinese	0.3%	0.4%

* number of months over 6 years at the end of academic year

Results of multiple regression and multilevel modelling

Table 5-12 and Table 5-13 present the results of all our models that estimated the ECC effect. Table 5-12 presents estimates using KS1 mathematics points as the outcome and Table 5-13 presents results from equivalent models using KS1 English points as an outcome. Results from both OLS multiple regression and multilevel modelling are presented.

Table 5-12: Estimates of the effect of ECC using KS1 mathematics points as an outcome

	Multiple Regression results				Multi-level Model results	
Model	ECC effect in KS points (SE)	ECC effect in SD units (SE)	Adjusted R-squared	Number of pupils	ECC effect in KS points (SE)	ECC effect in SD units (SE)
1	-0.265 (0.034)	-0.070 (0.009)	0.001	49521	-0.039 (0.094)	-0.010 (0.025)
2	0.045 (0.026)	0.012 (0.007)	0.435	47241	0.141 (0.071)	0.037 (0.019)
3	0.032 (0.025)	0.008 (0.007)	0.481	46656	0.130 (0.067)	0.034 (0.018)
4	-0.034 (0.025)	-0.009 (0.007)	0.487	46656	0.110 (0.066)	0.029 (0.017)

Table 5-13: Estimates of the effect of ECC using KS1 English points as an outcome

	Multiple Regression results				Multi-level Model results	
Model	ECC effect in KS points (SE)	ECC effect in SD units (SE)	Adjusted R-squared	Number of pupils	ECC effect in KS points (SE)	ECC effect in SD units (SE)
1	-0.417 (0.037)	-0.101 (0.009)	0.003	49527	-0.231 (0.100)	-0.056 (0.024)
2	-0.075 (0.027)	-0.018 (0.007)	0.451	47235	-0.016 (0.079)	-0.004 (0.019)
3	-0.088 (0.026)	-0.021 (0.006)	0.532	46650	-0.026 (0.072)	-0.006 (0.017)
4	-0.163 (0.026)	-0.039 (0.006)	0.539	46650	-0.056 (0.070)	-0.014 (0.017)

Figure 5-10: Estimates of the effect of ECC using KS1 mathematics points (in SD units) as an outcome

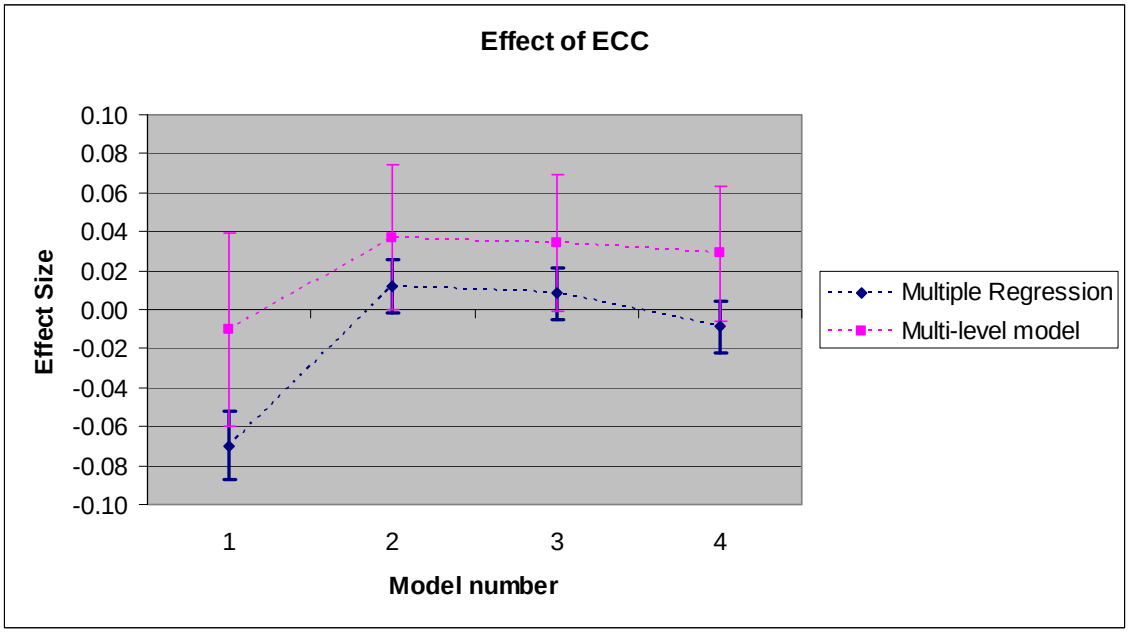
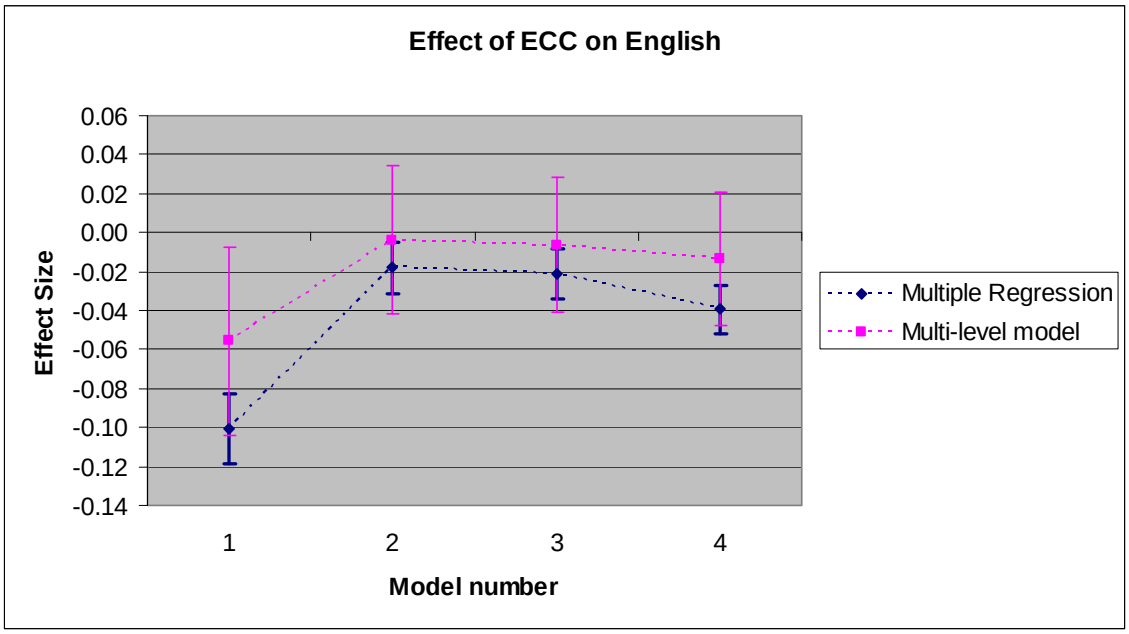


Figure 5-11: Estimates of the effect of ECC using KS1 English points (in SD units) as an outcome



The results in Table 5-12 show the ECC effect using KS1 mathematics points as an outcome and multiple regression. The Null model (1) shows a statistically significant effect of -0.265 in KS1 mathematics points. However we know from the descriptive statistics that the intervention group had lower prior attainment scores compared to the comparison group. When we controlled for these differences in model 2 the ECC effect dropped to almost zero and was not statistically significant at the 5% level. Similarly models 3 and 4 failed to detect a significant ECC effect. The multilevel models produced slightly more positive effects but once prior attainment was controlled for the effects were still not statistically significant. These results are illustrated more clearly in Figure 5-10.

The results of the four models with KS1 English points as an outcome are presented in Table 5-13. Model 1 shows that the intervention group had lower KS1 English results compared to the control group. Once prior attainment was controlled for, models 2 to 4 showed small, statistically significant, negative effects. We can see this clearly in Figure 5-11. This appears to show a small negative ECC effect on the pupils KS1 English results. However the multilevel models again showed slightly more positive effects although, apart from the null model, they were not statically significant. It is perhaps worth noting that multilevel modelling accounts for variation at school level as well as pupil level, and that differences between schools have an impact on the results.

Subgroup analyses

As we did not detect a main effect it was inappropriate to undertake any sub-group analyses.

5.4: Discussion and conclusions

We conclude that the results of the interrupted time series design and the case control designs were equivocal. Overall using these four quasi-experiments we were unable to detect a medium-term effect on mathematics that could be attributed to ECC. Using a quasi-experimental approach no evidence of an effect was found. Quasi-experimental designs are less robust than RCT designs, and although we controlled for as many differences as possible between the control and comparison groups, there will always be unknown or immeasurable differences that cannot be accounted for. RCT designs, on the other hand, control for these unexplained differences by virtue of the design.

Chapter 6: Economic evaluation

Key summary points

- Key findings are based on figures provided by Department for Education (DfE) which showed that it cost on average £1353 for each child to receive the programme.
- We found that Numbers Count led to an extra 9% of children working at the equivalent of key stage 1 (KS1) level 2c immediately after receiving the intervention in January, which was 5 or 6 months before the normal assessment time at KS1.
- Numbers Count delivered in pairs is more cost-effective than Numbers Count delivered as an individual programme.
- The cost per additional child working at the equivalent of a KS1 2c or above on the Progress in Mathematics 6 (PIM 6) is approximately £5000 for pairs compared with usual teaching and £15,000 for one-to-one compared with usual teaching.
- The cost per extra numeracy week gained by the intervention is approximately £193 for each child for the individual intervention.

6.1: Introduction

We examined the cost-effectiveness of Numbers Count (NC) and adapted NC delivered to small groups of children (pairs and triplets) using outcome data from Trials 1 and 2.

In addition to understanding the effectiveness of an intervention, it is very important to investigate its cost-effectiveness. In a cost-effectiveness analysis (CEA) the benefits are described in their natural units – in this case, within Trials 1 and 2, this is the cost per additional child achieving the equivalent of a level 2c or above at key stage 1 mathematics on the PIM 6 mathematics test (as measured in January or April, up to five or six months before normal testing using this measure). Because we are using a CEA we can only say Every Child Counts (ECC) is cost-effective if the intervention generates more benefit at lower cost than the alternative. For ECC (receiving NC) compared with normal classroom teaching (not receiving NC) this will not be the case as our evaluation is restricted to the timeframe of the short-term impact of NC as evaluated in the randomized controlled trials (Trials 1 and 2), so any future possible cost savings that might outweigh the initial cost of ECC remain unidentified.

This evaluation assesses whether receiving NC or adapted NC for groups is the more cost-effective policy when compared with not receiving NC. The trials were described in Chapters 2, 3 and 4 of this report and essentially comprise 3 comparisons of different modes of receiving NC, compared with not receiving NC. The trials assessed the effectiveness of:

- Normal classroom teaching in mathematics
- Normal classroom teaching in mathematics plus NC delivered to individual children

- Normal classroom teaching in mathematics plus adapted NC delivered to pairs of children
- Normal classroom teaching in mathematics plus adapted NC delivered to triplets of children

Trial-based evaluations were conducted for each of these comparisons, to assess the relative effectiveness of different modes of delivery of NC.

6.2: Methods

The economic evaluation based on the trials reported earlier addresses the following question: What is the cost-effectiveness and incremental cost-effectiveness of the three types of NC delivery compared to normal classroom teaching (no NC) and to each other? Our cost-effectiveness results only apply to a single term. Because of the lack of a long term comparator group we cannot estimate whether or not the effectiveness of NC or adapted NC is sustained or whether it could lead to future resource savings.

We undertook 3 comparisons in the economic evaluation:

- What is the cost-effectiveness of receiving NC compared with not receiving NC?
- What is the cost-effectiveness of adapted NC delivered in pairs compared with NC delivered to individual children?
- What is the cost-effectiveness of adapted NC delivered in triplets compared with NC delivered to individual children?

We estimated the costs of NC and adapted NC delivered to pairs or small groups of children (triplets) to enable us to calculate the costs of each mode of delivery. Outcomes from each mode of delivery are based on those previously reported. We compared these costs with the outcomes from the programmes to assess the incremental cost per additional child working at the equivalent of Level 2c or above at key stage 1 mathematics on the PIM 6 mathematics test (as assessed in January 2010 or April 2010). We estimated the costs of delivering NC based on Numbers Count teachers' (NCTs') salaries (DfE, *personal communication*, 2010). However, some lessons were delivered by deputy head teachers or teacher leaders (TLs). Therefore our estimates may tend to underestimate the costs of delivering NC.

6.2.1: Primary outcome

The primary outcome measure, from the randomized controlled trials of the NC interventions, was achievement on PIM 6 in January 2010 (and April 2010 for Trial 2), measured for all three comparisons in the evaluation. We assessed the extrapolated cost per child working at the equivalent of level 2c or above at key stage 1, estimated from achievement on PIM 6. We show the conversion scores in Table 6-1.

Table 6-1: Estimates of National Curriculum level (and sublevel) associated with PIM 6 raw scores, from p.41 Progress in Mathematics 6 Teacher's Guide

PIM 6 Raw Score	Estimated National Curriculum Mathematics Level
0-9	W - working towards level 1
10-13	1c
14-17	1b
18-20	1a
21-23	2c
24-26	2b
27-28	2a

GL Assessments, which developed PIM 6, published predicted estimated levels for National Curriculum Mathematics associated with the raw test scores (Table 6-1). These scores show the estimated *current* level that a child would be working at under the National Curriculum given their PIM 6 score. The assumption was that children in this evaluation would do at least as well as this when they were assessed at the end of key stage 1 during the summer term of the same academic year. This sets a lower bound for the analysis.

All future costs have been adjusted to 2009 prices (using a discount rate of 3.5%, as recommended by HM Treasury, Green Book) and presented in both discounted and undiscounted form. Discounting is where future costs are converted into present day values. Costs occurring in the future have less value than those occurring at the present time, which needs to be reflected in the cost calculation. The sunk costs of developing the NC intervention (initial costs) are not considered in the evaluation since they are not affected by programme roll out, so whether or not the programme is implemented in the future will not lead to any savings of these costs.

Cost per child and the additional probability that a child will achieve a score of 21 or above on the PIM 6 has been presented for each option (normal classroom teaching without NC, NC, adapted NC in pairs, and adapted NC in triplets) and linked in ascending order of costs (from least expensive to most expensive). Dominant options were identified, and for non-dominant options we calculated appropriate incremental cost-effectiveness ratios (ICERs). We assumed that total costs for a roll out of the programme nationally would be a sum of individual costs multiplied by the number of children offered the programme. Therefore, we did not need to estimate the total number of children nationally who would be eligible for this programme.

6.2.2: Costs

Resource use is based on the direct costs to the education sector, which includes the DfE, local authorities, and schools. Resources used outside this sector are excluded, for example any costs that fall on pupils, their families, other sectors and any productivity changes, as well as capital costs (Drummond et al., 2005).

Our primary sources of costs are from a report by KPMG, which estimated the cost components of the ECC initiative (Every Child a Chance Trust, 2009). In addition, we were provided with a breakdown of salary costs that the DfE actually used for ECC. For our primary analysis we used a combination of KPMG costs and the salary costs provided by the DfE. This combination produced a lower cost estimate than using the KPMG costs alone.

However, in a sensitivity analysis, we also included an analysis using KPMG costs alone, updated to 2010 prices. We did not have any information regarding other costs that might be recovered in the future, such as less teaching time to support ECC children in the future.

6.2.3: Synthesis

We combined the costs and outcomes using an ICER. If one of the NC interventions is more effective and costs less then it is said to 'dominate' the alternative intervention. For example, if group teaching produces higher mean mathematics scores and at a lower cost than individual teaching then it is said to be the dominant intervention. However, a situation that is quite common is for the more expensive intervention to be more effective than the less expensive alternative. In this case this is the cost per additional child in the NC group getting higher than the mean of the control group. This is the incremental cost-effectiveness ratio (ICER), which is the difference in costs and impact of two programmes (Drummond et al 2005).

6.3: Results

6.3.1: Costs

In Table 6-2 we show the cost estimates we used for the economic evaluation broken down by category. The costs for the economic evaluation were sourced from a combination of DfE self-reported provided financing (DfE, *personal communication*, 2010) and publicly available costs (Every Child a Chance Trust, 2009).

Table 6-2: Costs of ECC

Costs	Value	Source	Adjusted to 2010(Q5) using RPI
Average annual DfE financing for an existing NC teacher (0.5) for 1-1 teaching	£13,589	DfE	
Calculated over 4 years (undiscounted)	*£54,356*		
Cost of training course to school	£4,500.00	KPMG	
Extra mathematics resources	£1,000.00	DfE/KPMG	
Total average school costs over 4 years	£59,856		
Average annual school costs (Total cost/4)	**£14,964**		**£15,368**
Average Local Authority annual costs per school (net)	**£826**	KPMG	**£849**
Total combined school & Local Authority annual net costs	**£15,790**		**£16,217**

We assumed the total cost for the programme was fixed and consequently the cost per child depended upon the number of children that received NC and adapted NC programmes. Table 6-3 shows the cost per child for the different teaching programmes.

Table 6-3: Annual cost per child by type of NC programme

	Number of children taught annually	Annual additional cost per child (2008 prices)	Annual additional cost per child (2010 Q5 prices)
Usual Teaching	0	£0	£0
1 to 1	12	£1,314	£1,353
2 to 1	18	£877	£902
3 to 1	27	£585	£601

The following assumptions underpin the values in this table. First, children received 30 minutes of teaching per day for a maximum of 12 weeks in one term so that on average children received 30 hours in the year. Second, classes were not missed and teachers and students were not absent (this is a conservative assumption and will underestimate the costs; thus it is likely that the true costs are higher than we estimate). Third, for individual teaching, a half time teacher taught four children per term in four available slots. Fourth, for pairs and triplets delivery, teachers only had three slots available for teaching, with the fourth slot taken up with extra administration associated with teaching the additional children. The process evaluation noted that in some cases children receiving the adapted NC in pairs and triplets received up to 45 minutes of support. However, our analysis strategy was by intention to treat, so we did not make any adjustments for varying lengths of time in adapted NC. This was also in line with our pragmatic approach in Trials 1 and 2, and had the effect of increasing the impact of adapted NC. The reduced number of teaching sessions was the format used in the trials and it implies that somewhat less than double or triple the number of children can be taught using adapted NC for pairs and triplets. This assumption about the potential extra children who can be taught overestimates the costs of pairs and triplets in comparison to individual teaching. The assumption was also made that teachers were in their second year of teaching, but teaching adapted NC for the first time.

In Table 6-4 we show the proportion of children who would be potentially working at the equivalent of level 2 or above at KS1 mathematics as estimated by the PIM 6 mathematics test (assessed in January 2010 or April 2010). As the table shows, both NC and adapted NC in pairs and triplets were more effective than normal classroom teaching.

Table 6-4: Impact of the intervention on proportion of children working at the equivalent of level 2 or above at key stage 1 mathematics

Impact of intervention: PIM 6 comparisons cut-score 21 from ECC trials	**Number**	**Proportion**	**SE**	**Lower 95%**	**Upper 95%**
All trials: one-to-one vs. control					
Usual teaching	41/434	9%	0.014	7%	13%
Intervention	30/163	18%	0.030	13%	25%
Trial 2: one-to-two vs. usual teaching					
Usual teaching	9/100	9%	0.029	4%	16%
Intervention	14/53	26%	0.061	15%	40%
Trial 2: 3:1 vs. control					
Usual teaching	9/60	15%	0.046	7%	27%
Intervention	12/39	31%	0.074	17%	48%
one-to-two vs. one-to-one					
NC one-to-one	3/25	12%	0.065	3%	31%
NC one-to-two	26/104	25%	0.042	17%	34%

Note: The combined one-to-one versus one-to-two comparison shows superiority for one-to-two in this analysis which differs from the main trial analysis, which used the PIM 6 as a continuous variable not a binary one as here. In addition, the very small sample sizes lead to large uncertainty (e.g., the 12% of children getting 21 score or above for the one-to-one comparison against the one-to-two of 25% is based on only 3 out of 25 children).

In Table 6-5 we show the incremental cost-effectiveness ratios for each NC intervention. Compared with normal classroom practice the interventions were more expensive yet produced more benefit. For NC compared with normal classroom teaching (no NC) the discounted cost per extra child reaching the equivalent of level 2c or above at KS1

mathematics was approximately £14,600. For the pairs adapted intervention the ICER was around a third of this, partly due to the lower cost but also partly due to the slightly increased effectiveness of NC delivered in pairs demonstrated in Trial 1 (autumn term), although this was not statistically significant (a cost per additional child of approximately £5000).

Table 6-5: Incremental cost-effectiveness ratios

2010 Q5 prices	**ICER per year**	Lower 95% for annual	Upper 95% for annual
one-to-one vs. usual teaching			
Undiscounted	**£14,611**	£10,369	£22,143
Discounted	**£13,656**	£9,691	£20,695
one-to-two vs. usual teaching			
Undiscounted	**£5,010**	£3,646	£7,890
Discounted	**£4,683**	£3,408	£7,375
3:1 vs. usual teaching			
Undiscounted	**£3,689**	£2,771	£5,862
Discounted	**£3,448**	£2,590	£5,478
one-to-two vs. one-to-one			
Undiscounted	Dominated (less costly and more effective)		
Discounted			

Note: although the cost per child for one-to-two is only about a third less expensive than one-to-one the cost-effectiveness ratio appears a lot lower because the difference in proportions scoring 21 or above are somewhat greater and this will lower the cost-effectiveness ratios.

Earlier in the report (Chapter 2) we estimated in Trial 1 that NC produced on average an additional improvement of 7 weeks in numeracy skills, as measured by the PIM 6, compared with no NC. In other words, in a 12 week term the NC children improved by 19 weeks compared with 12 weeks' improvement of the control children. In terms of cost-effectiveness, this equates to a cost of approximately £193 per additional week of progress per child.

6.4: Discussion

This economic evaluation was based on estimates of costs and effects within a single term. One potential scenario is that the intervention costs about £14,600 to help one child work at the equivalent of level 2c at key stage 1 mathematics (estimated in January) with this differential effect being lost with the control children catching up. Alternatively, if we assume

that the average child in the NC group made an additional seven weeks' progress compared with the average child in the control group then the cost of achieving each additional week's progress for one child was approximately £193.

Adapted NC looks relatively cost-effective compared with NC. In this analysis we found that more children receiving adapted pairs NC were working at the equivalent of level 2c or above at KS1 mathematics as estimated by the PIM 6 test than were those children randomized to individual NC at a lower cost. This result differs somewhat from the result in the main effectiveness analysis where the PIM 6 is treated as a continuous variable. However, when the data are treated as continuous there is hardly any difference between the two groups in terms of effectiveness, and given that the one-to-two adapted intervention is significantly less expensive, it remains the dominant intervention and should be chosen in preference to one-to-one NC. Therefore, if there is a choice between NC and adapted NC delivered in pairs it would seem sensible to recommend that adapted NC should be adopted as the more cost-effective alternative. With respect to delivering NC in triplets of children - there is less certainty here due to the very small numbers of children in the feasibility trial which evaluated NC compared with adapted NC delivered to triplets of children (Trial 2: Triplets). Delivering adapted NC in triplets is less expensive and therefore probably should be preferred to the one-to-two adapted intervention. However, due to lack of numbers of children allocated and the extent to which the confidence intervals overlap in the analysis in this trial (Trial 2: Triplets) then we emphasize caution when recommending adapted NC in triplets.

Whether or not adapted NC in pairs should be delivered at all will depend upon the value given to one child working at the equivalent of 2c or above at key stage 1 mathematics because we have only been able to reliably assess the short-term impact - one term only. If it were considered to be worth around £5,000 or more then we should consider rolling out the programme.

In conclusion, adapted NC delivered in pairs is more cost-effective than NC; however whether it is more cost-effective than usual teaching with no NC in the long term would need to be reliably assessed by way of a long-term RCT.

Chapter 7: Process Evaluation: The effective implementation of Numbers Count

Key summary points

- The key findings are based on a cross-sectional study of a sample of participating Every Child Counts schools, which elicited views and perceptions of Numbers Count by key stakeholders.

- The Numbers Count teachers were all positive about NC, highlighting the impact on the children's mathematical and wider skills. All the NC lessons observed contained important lesson characteristics of building on previous knowledge, practise of fact retrieval, questioning, use of concrete and visual materials, and practical engagement of children. However, only about 25% contained problem solving tasks.

- The Numbers Count teachers rated their professional development very positively, highlighting in particular the opportunities to discuss practice with colleagues, and the support provided by teacher leaders and national trainers.

- The head teachers were also very positive about Numbers Count, the impact on the school and mathematics teaching. However, concerns were voiced about the long term sustainability of the existing model, including recruitment and retention of specialist teachers.

- The parents were all pleased about the perceived positive impact of the programme on their children. Generally, they valued the increased contact with the schools, in particular through the Numbers Count teacher. Many parents gave illustrative examples of how the programme had helped their children improve their mathematical skills, as well as their attitudes towards mathematics and schools more generally.

- The local authority officers felt that Every Child Counts was an effective programme, which helped meet the needs of the participating children, and contributed to improved mathematics teaching throughout the schools. They too expressed concerns about the long term viability of the model, particularly given the perceived future financial climate.

7.1: Introduction

This chapter of the evaluation complements the impact assessments from the randomized controlled trials (RCTs) by using a qualitative research design (process evaluation) to evaluate the effective implementation of Numbers Count (NC). We identified and focused on key implementation features of the programme at the classroom, school, and local authority (LA) levels, as well as the training, infrastructure and support mechanisms. We report the views and perceptions of key stakeholders in the development phase of the programme, and researcher observations of the programme in practice.

7.2: Design and methods

We used a cross-sectional design to elicit views and perceptions of a sample of key stakeholders involved in the delivery of NC in schools, and supplemented these views with researcher observation of a sample of NC lessons. Such a design does not, however, allow strong causal claims to be made, and alternative explanations for our findings cannot necessarily be ruled out, for example, the impact of other policy changes.

Our methods included visits to a representative sample of 20 schools during the development phase of the programme to observe NC lessons and to interview Numbers Count teachers (NCTs) and head teachers in those schools. We also carried out interviews with local authority officers and the programme managers, and observed professional development sessions. The interviews were semi-structured, and were piloted in schools not taking part in the research. We are aware that, during some of the later interviews, there was a perception by some stakeholders that the programme was likely to be discontinued, and therefore the interview responses might have been different had we conducted them earlier.

The survey instruments (24 in total), response sheets, and the protocol outlining in detail the plan and procedures for the process evaluation were approved by the DfE and the Steering Group (see process evaluation appendices in Torgerson et al, 2011c). All of the interviewees gave permission for their interviews to be recorded on the basis that the recordings would only be used by the researcher carrying out the interview. These recordings and the researchers' handwritten notes were used to analyse the responses. In all cases interviewees were invited to make contact with the researchers if they wanted to discuss, add or clarify any of their responses.

Two online surveys with Numbers Count teachers (NCTs) were also carried out, in November 2009 and in July 2010. The surveys were emailed by Edge Hill University to 480 NCTs that they had email addresses for (75% of the total number of NCTs). 242 responded to the first survey and 262 responded to the second survey. Therefore the response rate was about 50% for the first survey and about 55% for the second survey. The general aims of the first survey were to (a) find out how teachers felt with regard to preparation for the important aspects of the NC role, (b) obtain teachers' views on how well they thought their NC lessons were working and what the possible areas were for improvement, and (c) find out how they viewed their professional development. In developing the questions to ask in the survey, we considered the four strands included in the professional development sessions (an opportunity to review progress and practice, a specific focus on some aspect of pedagogy, a specific focus on some theoretical aspect of learning mathematics, and a specific focus on some aspect of the wider role of the teacher), and the interview questions that we would ask teachers to probe for their opinions on the four strands, to enable the survey to inform observations of professional development sessions. We considered the roles of the NCT as outlined in the NC handbook from Edge Hill University, and the aspects of the NC lessons that would be observed. The second online survey provided an opportunity for confirmation of the important issues emerging from the NC lessons and the training of NCTs. The general aims of the second survey were to (a) confirm teachers' views regarding their professional development and their role, and (b) confirm teachers' views on how well they thought NC lessons were working, including their views on specific areas for improvement.

7.2.1: Study population

The study population comprised the following:

- Schools opting to participate in the ECC programme (NC, adapted NC in pairs and adapted NC in groups of three). This included NCTs, head teachers, class teachers, and other staff (n = 20).
- NCT and TL training events (n = 18).
- Parents, children and staff from the above (n = 8) face to face, and (n = 5) by telephone.
- Schools choosing not to opt in (head teachers only) (n = 6).
- Local authority officers (n = 12).
- ECC programme management, including Edge Hill University. Ongoing contact and feedback was maintained throughout the research period.

Hereafter references to all stakeholders - LA officers, head teachers, teacher leaders, national trainers, NC teachers, year 2 teachers, teaching assistants and parents refer to samples of the above. In addition, such references may also include data from Trial 1 and Trial 2 schools following further process evaluation data collection during the testing period.

The evidence gathering was carried out on the basis of two interrelated strands, the *learning* process and the *organisational* process. The learning strand investigated the delivery of NC and how it worked at the classroom level, professional development and on-going support of NCTs and TLs. The organisational strand looked at impact, role of schools and LAs and the views of parents.

For the learning element, the primary research method consisted of visits to 11 schools delivering NC, and observations of 23 NC lessons at these 11 schools. During the visits, interviews were carried out with NC teachers, pupils and Year 2 teachers and teaching assistants. A further 11 lessons involving adapted NC delivered to pairs or triplets of pupils were also observed. In addition, we attended 18 training sessions for NCTs and TLs, observing the sessions and interviewing the participants and the presenters. The NCTs' views of their teaching, wider aspects of their role and their views on their professional development were gained through two online surveys.

For the organisational element the primary research method consisted of interviews (face to face, telephone and group) with head teachers (n = 31), local authority officers (n = 12), and parents (n = 8), as well as ongoing contact with the programme managers from the Every Child a Chance Trust and Edge Hill University. We used a semi-structured interview format which provided both a common framework with which to assess key questions, and a degree of flexibility to explore important issues that arose during the interviews.

7.2.2: Selection of schools

Including a representative sample of schools was an important element in the design. We wanted, as far as possible, to avoid selection bias, which could have occurred if, for example, we had concentrated our efforts on those schools where the programme was perceived to be running very well.

The majority of the process evaluation of NC (individual) was undertaken in 11 schools from six different LAs in England. The LAs were identified by the ECC evaluation Steering Group. None was involved in either of the trials. From the entire list of participating schools three schools in each of the 6 LAs were randomly selected to be approached to take part in the research. However, one authority was withdrawn from the evaluation by the DfE, and another authority was approached to be involved in the evaluation of adapted NC in pairs. Three schools from this authority agreed to participate. In addition, the authority involved in the evaluation of adapted NC delivered to triplets of children was approached, and two schools agreed to participate. There were a variety of reasons why schools declined or were unable to take part, for example imminent changes in staff and OFSTED inspections. Nonetheless, we are confident that this was a representative sample of those schools which had agreed to take part in the development phase of ECC. During the testing for Trials 1 and 2 members of the research team visited 12 intervention schools not involved in the process evaluation; we were able to carry out interviews and observations in these schools to corroborate the findings from the original sample. The learning and organisational strands overlapped at the school level. Visits to schools and interviews with teachers were carried out between September 2009 and May 2010. Observations were based on the identification of desirable characteristics of lessons with children facing difficulties in mathematics, and highlighted in the literature. Specifically, the desirable characteristics that the researchers looked for in the Numbers Count lessons were:

Structural characteristics: a focus on specific mathematical difficulties; building on existing knowledge of children and consolidation of earlier learning; practise of fact retrieval; incorporating problem solving tasks.

Pedagogical characteristics: use of questioning for assessing understanding; use of concrete and visual materials; practical engagement of children in activities; opportunities for discussion, explanation from children and reflection on methods from children; opportunities to develop correct mathematical language.

Interviews with NCTs and Year 2 teachers and teaching assistants focused on identifying the initial difficulties faced by children, the progress made by children in the NC lessons, positive characteristics of the intervention and areas for development, how schools maintained children's progress in mathematics, wider benefits of NC, and areas where teachers would like further support.

7.2.3: Analysis

The data from interviews were analysed for themes emerging from the data, making connections between issues highlighted in observations and also the results of the online surveys of NCTs. Our general principle was to only report findings (or respondents' views) where there was triangulation through multiple observations. A systematic process was used to analyse the raw data. The responses on the sheets were checked and compared with the recording to ensure accuracy and consistency, and where appropriate, to enable amendment or to add responses. The response sheets were then coded in terms of the research objectives and detailed questions. These coded responses were then used to generate the descriptive statements for the report. The key advantages of this approach were that it allowed the complex nature the ECC programme to be explored and understood, particularly in terms of its place in the wider educational context; it allowed unpredictable, but

relevant, issues and factors to be included; and it facilitated meaningful exploratory discussions about possible future options for the programme.

7.3: Results

7.3.1: The Numbers Count lessons

Valuable characteristics of the lessons

All 23 NC lessons observed contained the features identified through the literature: a specific focus, building on previous knowledge, the practise of fact retrieval, use of questioning, use of concrete and visual materials, and practical engagement of children.

The pupils were unanimous in stating their enjoyment of the NC lessons, mentioning particularly the games and activities. NCTs and Year 2 class teachers and teaching assistants were very positive about NC, mentioning particularly the key aspects of the diagnostic nature of the intervention and the one-to-one working with pupils.

Impact of the Numbers Count lessons

NCTs identified the main difficulties for children coming on to the programme were with regard to counting and the depth of their understanding. For example, they noted that some pupils understood the basics of a particular topic, but did not have a secure enough knowledge to undertake problem solving.

NCTs and Year 2 class teachers and teaching assistants stated that the NC lessons enabled them to identify pupils' progress in the areas of difficulties highlighted above, namely understanding of mathematics, confidence in mathematics and language, and this progress transferred over to the whole class setting.

Confidence of the Numbers Count teachers in delivering the lessons

The results from the first online survey found that NCTs were most confident in: setting up the lessons (87% of respondents replying *very confident*); setting up the NC area (72%); engaging children (70%); and using a variety of resources during the lesson (64%). However, NCTs were least confident about: incorporating *using and applying* into lessons (19%); using *reflection on next steps* in lessons (21%); *encouraging explanation and reasoning* (24%); and *diagnosing understanding* (27%).

Despite the *further applications opportunities* part of the lesson, only about a quarter of the 23 lessons observed contained opportunities to solve problems. Some of the teachers were concerned about how to cover this aspect of the lesson, given the constraints in the structure and timing of the 30 minutes lesson.

Suggestions for developing the Numbers Count lessons

From the observations and discussions with NCTs, a suggested area to explore in terms of developing the NC lessons from our perspective would be to remove one of the activities within the structure, in order to free some time. For example, the lesson could contain a

further application opportunity or a second *current learning activity*, rather than both. Examining this issue in the second survey of NCTs, a small majority (55% n = 144) agreed that the structure of the lessons was fine, but that the *further applications* part of the lesson could be removed and *using and applying activities* incorporated in other parts of the lesson. Only 28% (n = 73) agreed that the second *current learning activity* could be removed.

7.3.2: Wider role of the Numbers Count teacher

In addition to the delivery of the NC lessons, the results from the interviews highlighted the wider role of the NCT. Year 2 teachers and teaching assistants stressed the benefits that they had gained from liaising with the NCT, for example in tracking pupils' progress, or the NCTs being a source of ideas and resources for teaching mathematics. A theme that emerged from the interviews with teachers was that NCTs felt that their role was beneficial to other staff as it enabled them to share the expertise they had developed from the NC training and their experience of dealing with specific mathematics difficulties.

In terms of the support that NCTs were receiving in school, a further theme that emerged from the interviews was that almost all sample teachers were positive about the support they had received from head teachers and other staff. However, a number of NCTs expressed concerns that there was a lack of understanding amongst staff about the NC role. In terms of barriers to the wider NCT role, the first survey of NCTs found that almost a quarter of them had had no formal opportunities to liaise with colleagues in school, probably due to time constraints. Year 2 teachers and teaching assistants identified the need for more time for discussions with the NCT. Many NCTs stated that a 0.5 or 0.6 position was insufficient for the broader role, and the issue of combining the NC role with another role was a frequently raised difficulty. The issue of the amount of paperwork associated with the role was also raised. However, the teachers realised that the paperwork is part of the initiative and aids planning, and that great efforts had been made over time by Edge Hill University to reduce the burden of paperwork. In order to tackle these difficulties, some teachers stated that they would like more input during the professional development, possibly by bringing a colleague to the training. Teachers suggested that more appropriate information should be provided to schools, and more time should be allocated to the NC role. In order to support the progress of pupils after the NC intervention, 77% (n = 201) of respondents in the second survey of NCTs stated that they felt that very valuable suggestions would be more liaison with and information and training for teaching assistants, and 71% (n = 181) felt that more opportunities to liaise with the Year 2 teachers would be very valuable.

Teachers stressed the importance of parental involvement and its link with pupil progress in the intervention. Year 2 teachers and teaching assistants stressed the possible benefits of the programme on parents, in particular in terms of confidence in working on mathematics with their children. However, the first survey to NCTs found that the majority (69% n = 167) thought less than a quarter of parents of NC pupils had observed lessons; however, the same number also thought that parents were more involved in mathematical activities which were sent home with the children. Barriers to the involvement of parents included work commitments, caring for younger siblings, language difficulties (speaking a language other than English) and a general feeling of intimidation (including an awareness of their own difficulties with mathematics).

7.3.3: Professional development of Numbers Count teachers

Description of professional development

The professional development provided by Edge Hill University for NCTs consisted, in the first instance, of a series of one-day training days usually run by teacher leaders with local authority groups of NCTs (10 training days in the first year and 5 in the second year). Each training session contained the following elements: an opportunity to review, evaluate, discuss and share progress and practice in NC lessons; a specific focus on some aspect of pedagogy with the aim of developing NCTs' understanding of the teaching and learning of early mathematics; a specific focus on some theoretical aspect of learning mathematics, with the aim of developing NCTs' understanding of the learning of early mathematics; a specific focus on some aspect of the wider role of the teacher. Other professional development opportunities provided for NCTs were: teacher leader support to each NCT (5 in the first year and one visit in the second year) to observe a lesson and discuss any issues. Teachers were also encouraged to make one visit per term to a colleague or learning partner.

Teacher leaders were also provided with a professional development programme. These consisted of a national conference held at Edge Hill University at the start of the year, one or two other national events during the year and three area events for teacher leaders.

Positive aspects of the professional development and areas to develop

We attended 18 training sessions and interviewed two groups of three teachers per training session and one or two trainers per session during the professional development sessions. We found that teachers perceived the professional development positively. From observations, interviews with NCTs during the training, and the surveys of NCTs, it emerged that the most valuable aspect of the NC professional development was thought to be discussion between colleagues during these sessions, facilitated by the experience that the trainer brought to the sessions. The first survey and interviews with NCTs highlighted the support that the teacher leaders and the national trainers provided, and the value of the visits that NCTs received from teacher leaders. However, the opportunity to examine each others' practice through video during training sessions was seen as less useful than other aspects of the training. We observed that, although we thought the video was used very well in the training sessions in demonstrating aspects of NC practice, to illustrate theoretical issues and to share ideas, many teachers spoke of their difficulties in videoing lessons, finding time to watch their videos and edit clips. In the training sessions, deeper reflection on practice using the video was rarely observed. We recommend, therefore, that further support is provided to teachers on using the video as part of their professional development.

7.3.4: Pairs and small group teaching

The findings are based on 9 schools implementing adapted NC in pairs and small groups, as well as those implementing NC individually, some of which worked with pairs particularly towards the end of the term, and therefore provided a useful and meaningful additional perspective of the key issues. The 9 schools were not randomly selected. This was a light touch evaluation because we were unable to see both the intervention and re-integration in all cases, some classroom observations were carried out by another researcher, and

additional interviews with heads were carried out in schools where we had not observed the teaching. Altogether, 11 small group lessons were observed (8 pairs lessons and 3 triplets lessons) and provided some insight into whether these alternative approaches to lessons might provide additional benefits or contain particular drawbacks for the teachers and pupils.

Lesson and classroom observations

A number of potential benefits of pairs and small group teaching were identified from the observations: pupils listening to others' answers and explanations; learning to take turns and to work with others; good preparation for whole class work; other pupils stepping in to help pupils with answers; pupils providing explanations of why another pupil might have got the wrong answer; pupils telling others what the words said; showing other pupils' work to reinforce ideas; highlighting one pupil's misconceptions to others; opportunities for conversations between pupils and to pick up on interesting comments and insights into possible misconceptions; other pupils extending ideas.

A number of potential drawbacks to pairs and small group teaching were identified from the observations: more distractions for pupils and problems of keeping the attention of everyone; differentiation required by the teacher; some pupils dominating in answering questions; pupils relying on help from others; problems with the combinations of pupils.

The 11 pairs/triplets lessons observed tended to be longer than standard NC lessons (45 minutes).

NCTs' views regarding pairs and small group teaching were mixed, with some noting the interaction between children as being very beneficial for learning, but others pointing out the need to properly match the children in order to make the arrangement effective. In addition, NCTs raised the issue of problems arising because children being taught in pairs or triplets were absent, or because pupils were already working with lots of other staff around the school.

Organisational aspects of pairs and small group teaching

Pairs and small group working was viewed by most head teachers as more difficult to manage than teaching children individually, in terms of space and resources, and because of the nature of some of the pairings. The pairings had been determined through a combination of randomization and teacher decision and in some instances this led to pairs that schools thought were not ideal. It should be noted that if small group adapted NC were to be adopted in the future schools would be completely free to match pairs of children. Examples of the organisational difficulties of sub-optimal pairings mentioned included different levels of mathematical performance, specific differences in understanding as well as learning, and individual pupils missing lessons. Whilst most head teachers said they would be in favour of more pairs working where they could choose the makeup of the pair, they also said they would want to have the option of one-to-one working where they felt it was appropriate. It was strongly suggested that the initial assessment / diagnostic phase should remain a one-to-one activity.

The 12 LA officers were very supportive of the notion of pairs and small group working, and in particular emphasised the need to explore this further. Head teachers and LA officers did

point out that whatever arrangements were adopted it was essential that the fidelity to the programme was maintained. The parents interviewed had little objection to pairs working, although some were concerned about the possibility of their child losing the relationship they had built up with the NCT, but they mostly saw the value of developing (and learning to develop) new friendships, which for some children had been a problem.

The NCTs received a small amount of extra training for the small group work and taught fewer sessions (but had more children overall). Nonetheless it was pointed out that in effect teaching adapted NC in pairs was not simply *double ones*, but a different intervention, and the planning and resourcing needed to take account of this. Head teachers and LA officers believed that the main difference was that individual NC allowed the NCT to better identify and focus on specific individual needs, whereas they thought adapted NC to pairs could help children's' social issues.

7.3.5: The school context and adoption of Numbers Count

We visited a wide range of schools in many different contexts and circumstances and saw differing levels of knowledge and expertise with regard to NC and mathematics pedagogy. Whilst we saw evidence of formal and informal promotion of the programme, for example through publications such as newsletters, and direct contacts from TLs, we found occasional confusion about how the programme operated in some schools (although this was very rarely about the key aims themselves). For example, some schools took quite a narrow view of the programme, seeing it as a way of improving the mathematics performance of the target children, and not so much as supporting more fundamental changes to the teaching of mathematics in general. LA officers were generally well informed about all aspects of the programme. The importance of both LA officers and head teachers fully understanding the programme is likely to be greater as education provision, particularly intensive support, is commissioned on a more local basis.

The vast majority of the schools were in relatively deprived areas; a point made by some parents was that the schools did provide something of a haven of stability for them in their area. It was also mentioned that there was little history of parents being involved in their children's education, and relatively poor communication between school and home, a point echoed by many head teachers. However it was also pointed out that ECC had helped with home / school contact by encouraging parents to take an interest in their children's NC work.

Most head teachers had been in post for several years or more and felt well prepared for ECC. However several thought that they had not been made sufficiently aware of what was required, for example, because the school had only agreed to take part at a late stage. This point was supported by LA officers who talked about some of the challenges of recruiting schools that they thought would benefit from the programme.

There was a general consensus that to get the best from the programme a relatively collegiate style of management was most appropriate, and a few head teachers pointed out that a directive approach would not be very effective, but rather that NCTs needed to have a good degree of flexibility and autonomy. This view was supported by the LA officers. In essence it was recognised that NCTs had to be able to develop a relatively independent and entrepreneurial approach to the programme.

Quite a few of the schools were keen to take part in many of the pilot programmes and new initiatives on offer. Indeed this was confirmed by the LA officers, who, in some cases, described schools as a safe bet or an enthusiastic early adopter when it came to trialling new programmes.

Given the relationship between numeracy and literacy difficulties, a few schools also took part in the Every Child a Reader (ECaR) programme. Many of these were able to integrate their ECC and ECaR working effectively, and it was suggested that doing both helped to normalise the notion of intensive one-to-one interventions. Nonetheless a number of schools pointed to the difficulties of adopting more than one major initiative at a time. Issues raised included availability of suitable staff, cost, and initiative overload. This was felt to be the case for head teachers with teaching commitments, limited senior staff support and lack of suitable space. In two of the schools visited it was apparent that this had been a problem, although given the number and nature of the selection criteria for the research, it would be inappropriate to generalise too much from this finding.

A limitation frequently cited by head teachers (and LA officers) was that if they adopted ECC only, they thought that by definition there must be children with reading difficulties in their schools whose needs they would not be meeting. Several head teachers said that it would be unfair to choose between ECC or ECaR.

7.3.6: NC teacher recruitment and management

NCTs had a variety of backgrounds but were generally very experienced teachers and head teachers thought this was an essential pre-requisite for being appointed to the role. Most of the head teachers thought it was useful for NCTs to make connections and contacts in the local community. In terms of recruitment, concerns were expressed by many head teachers about the long term availability of suitable teachers. Indeed a number said they had been lucky to find the person they had recruited to the role. The lack of a career path and relative job insecurity was believed to have possibly discouraged many potential candidates. NCTs' contracts varied between 0.5 and 0.6 of a full-time position, with there being no clear explanation as to why this variation existed in different schools. Whilst most head teachers thought the appointments should be 0.6 FTE, the most important thing was consistency across all schools and clear guidelines on roles and expectations.

We received universal praise from the head teachers about the quality and suitability of the training and professional development the NCTs received. Many head teachers took an interest in the training, and reported that it was the best they had ever seen, and should serve as a benchmark for future training/professional development. There was, though, less enthusiasm for the MA option in terms of the relevance to NC teaching, although some head teachers recognised the value of the higher level qualification in terms of contributing to the development of mathematics teaching more generally. It was also pointed out that it could provide a career development support for NCTs which is important in term of recruitment.

NCTs were managed and supported both by the school and the LA through the TL, and we were not made aware of any significant difficulties of the arrangement. Head teachers reported that NCTs were treated very much as any other teaching staff, and as far as possible they took part in and supported, wider school activities. The amount of direct

management and contact from head teachers varied, and was largely determined the level of experience and expertise of the NCT, and the management ethos of the school.

7.3.7: Wider organisational issues

There was considerable variation in the roles and responsibilities of NCTs in the schools. In most schools the NCTs contributed to school wide activities such as providing mentoring to less experienced colleagues. We found that most head teachers said their NCT took an active and sometimes leading part in staff meetings. We were also given many examples of where the NCT had become the champion for some of the children they taught, and this role was particularly valued by parents. LA officers valued their input, and overall, the role and contribution of NCTs to the schools was generally highly valued by head teachers and in many instances by the parents too.

Most of the head teachers had been closely involved in the pupil selection process, and although selection criteria had been established (and published in the handbook) there was nonetheless considerable variation in how schools actually selected pupils. In general terms head teachers were happy with the criteria, but many thought they only established a starting point from which to consider the most suitable children (within the spirit of the guidance). Some schools carried out screening with objective tests for all the cohort (for example assessments such as Rising Stars, Progress in Mathematics, and Performance Indicators in Primary Schools), and typically the results were used as the starting point from which to select the children. Even in those schools that did not do this we nonetheless found quite strong support for the notion of objective screening to identify children with mathematics difficulties. In practice most children were selected using both teacher judgements and previous assessments. In a few instances NC was used as a reward or as incentive for children with behavioural problems, and some schools targeted children who they felt would make the most progress. All of the parents were happy that their children had been selected, and felt that the intervention was appropriate for their needs. There were difficulties in selecting the bottom 5% of children in a school, and it is unlikely that these decisions coincided with the bottom 5% nationally, which remains a major challenge for the programme.

Finding a suitable area for delivery of NC proved a challenge for a few schools. The required standards for the NC area were high, and although some head teachers had concerns about this, the overwhelming view was that this was right. The contact and involvement of parents was thought to be one of the main benefits of the programme. An explicit aim of ECC is to involve parents and carers in the NC process and most schools had achieved this aim. From a more general perspective, many of the schools said that ECC had helped them make contact and build up more of a working relationship with ‘hard to reach’ parents and families. Many of the schools had organised workshops and open days/evenings for the parents which were generally thought to be successful, although some parents were reluctant to take part, and more direct personal contact was needed.

The views of the schools were largely echoed by the parents. They valued and appreciated the contact. For instance the Numbers Count diary was felt to be a useful and valuable way of maintaining contact with the NCT and the school generally. Indeed a few parents pointed out that the positive and supportive as well as the more immediate nature of this communication was very useful and pleasant to receive. Several parents said that the

reports they normally received from the school were sometimes difficult to understand. For the vast majority of the parents the NCT became the key contact, and several parents of children who had finished the programme said they had maintained informal contact with the NCT after completion.

7.3.8: School and LA partnership

The most commonly stated reason for a school taking part in ECC was a perceived weakness in mathematics, particularly the key stage results. In many cases this was largely based on the ranking of the key stage tests, which may not necessarily indicate a weakness in mathematics teaching *per se*, but rather be related to the low starting point of the children. A number of head teachers recognised that the programme provided an opportunity to address some of these weaknesses. In a similar way LAs tended to target ECC at the lower ranking schools. However, it was pointed out that this might lead to low performing children in high performing school not receiving the programme.

In general terms most of the head teachers felt that the LAs, and the relevant documentation, provided a reasonable and fair picture of what would be required of them. However, a few schools found themselves committing more time than expected to setting up the programme and to routine activities such as the exit testing. Most of the head teachers were of the view that such additional costs were part of any new intervention, and once bedded down would not be such an issue. Furthermore, when set against the wider benefits, the overall effort was felt to be worthwhile. A number of head teachers felt more could be done to promote meetings between head teachers participating in ECC, and whilst this was done in some instances, there were opportunities to combine such meetings with existing routine LA / cluster meetings.

7.4: Conclusions

The overwhelming message to emerge from the process evaluation of the implementation of NC was the positive response to Numbers Count and ECC in terms of the programme, the training and support infrastructure. The programme was seen to be well implemented in almost all of the sample schools and LAs. This was thought to be mainly due to the training and support network provided by Edge Hill University and the national trainers. It should also be noted that we were evaluating the development phase of the programme, and relevant to the longer term, we received positive comments from NCTs and TLs, as well as schools and LAs about how they felt able to feedback their views and experiences to the programme managers. This is an important finding in view of the likely future of intensive support programmes such as NC, and in particular the need for such programmes to be able to adapt to future circumstances and challenges.

The process evaluation has a number of strengths. It allows the key stakeholders associated with NC at all levels to have a voice in the evaluation, and it allows the complex and interconnected nature of such an intervention to be more fully explored and understood than in the trials and quasi-experiments, which it complements. In addition it allows future options and scenarios to be explored, which may help generate possible solutions for the future challenges (in particular to address funding constraints). However, there are also a number of limitations. The design of the process evaluation is based on views and

perceptions of sample stakeholders. Although we based our findings on consensus views (unless otherwise stated) some of the stakeholders might have had conflicts of interest which might have influenced their views and perceptions of NC. In addition some schools that volunteered to participate in ECC may have been systematically different from schools not participating in ECC (for example better equipped to make a success of the programme).

Chapter 8: Lessons learnt and future challenges

Key summary points

- The key findings are based on a cross-sectional study of a sample of participating Every Child Counts schools, which elicited views and perceptions of Numbers Count by key stakeholders. The design does not allow causal inference; and other alternative explanations for our findings cannot be ruled out.
- The diagnostic, intensive nature of the intervention was highlighted as a particularly important feature of Numbers Count that could be applied to other circumstances. Looking more to the future, it was suggested that the Numbers Count structured diagnostic process was something that many schools could adopt.
- The high quality and value of the training and ongoing professional support was recognised as a particularly important lesson from Numbers Count. Whilst the associated costs of the training were relatively high compared to other professional training, it was felt by many Number Count Teachers and schools that it represented good value, and that Numbers Count professional development should set the standard for training in schools more generally.
- The opportunity to examine each others' practice through video during training sessions was seen as less useful than other aspects of the training. Teachers spoke of their difficulties in videoing lessons, finding time to watch their videos and edit clips. In the training sessions, deeper reflection on practice using the video was rarely observed.
- NC was thought to be an excellent catalyst to promote home school contacts and relationships. The key was felt to be the involvement of the parents as far as possible in the actual process of learning (rather than as a passive observer). This is a feature that could be adopted for other interventions.
- Many of the head teachers felt that the intensive approach taken by Numbers Count would lead to other wider benefits, in particular for the children on the programme. There was a widespread view that mathematics teaching in general in the school would improve with the presence of the highly qualified Numbers Count teachers. The key challenge for head teachers was sustainability, and, for many, a realistic view was that it was unlikely that Every Child Counts would continue in its current form.
- Local authority officers saw a strategic value in adopting a well planned and resourced programme such as Numbers Count in raising mathematics standards in weaker schools. However, for them, a key challenge was being able to manage the resources in such a way as to accurately target the bottom 5% nationally. There were concerns about the long term sustainability, and in some instances appropriateness, of such an expensive intervention.

8.1: Introduction

This chapter of the evaluation seeks to bring together the key lessons learnt from the process evaluation of the implementation of Numbers Count (NC) that could be applied to other situations, and identifies some of the key challenges which would be faced should the programme be rolled out.

8.2: Design and methods

As outlined in Chapter 7, we used a cross-sectional design to elicit views and perceptions of a sample of key stakeholders involved in the delivery of NC in schools, and to supplement these views with researcher observation of a sample of NC lessons. As pointed out previously, such a design does not allow strong causal statements to be made; and other alternative explanations for our findings cannot be ruled out. We draw on the data collected for the previous chapter; 'The effective implementation of Numbers Count', and relate these to wider mathematics teaching and school organisational issues.

8.3: Findings

8.3.1: Impact on mathematics pedagogy

All Numbers Count teachers (NCTs) and Year 2 teachers and teaching assistants interviewed were very positive about NC, commenting that they believed the pupils made good progress and increased their understanding as a result of the programme. There were two key interlinked reasons why the majority of the teachers considered this to be a successful programme leading to excellent progress. These were the detailed *diagnostic* nature of the intervention and the one-to-one working. Teachers felt that the detailed diagnostic work enabled them to specifically identify mathematical difficulties that would be missed in a classroom setting. A number of teachers wondered how many similar students have slipped through the net in the past, without having had access to such an intervention. The one-to-one working was considered to be crucial as it enabled the teacher to tailor everything to the individual pupil's needs and pace of working. Additionally, some respondents commented that they thought it was important that the intervention was delivered early on in the pupil's schooling so as to deal with difficulties before any gaps in understanding developed.

In her review of the literature of what works with children with mathematical difficulties, Dowker (2004) highlighted the finding that interventions that focus on the particular components with which an individual child has difficulty are likely to be most effective. She also noted that children without difficulties used a far greater range of strategies than pupils with difficulties, and mathematical language was also an area of difficulty for these pupils. Based on these difficulties, the literature provides a range of recommendations for teaching children with difficulties in mathematics. Dowker (2004) made the following recommendations for teaching children with difficulties: revising and consolidating earlier learning and rehearsal of earlier learning; provision of appropriate concrete and visual materials that can be used for the solving of problems; a variety of activities including multi-

sensory approaches – something to see, listen to and do; opportunities for discussion; highlighting and using number patterns; practise for fact retrieval and reasoning.

Haylock (1991) recommended the avoidance of reliance on routines and algorithms, disembedded tasks and purposeless activities, and moving children on too quickly to new topics. Anghileri (2001) also highlighted the problems associated with routines and algorithms, stating that they encourage "'cognitive passivity' and 'suspended understanding' because they do not correspond to the way people naturally think about numbers" (p. 25). Denvir & Brown (1986), in their research on working with low attaining 7 to 9 year olds, presented the following recommendations for teaching: children must be active, both in interacting with the physical world and in reflecting on these interactions; ideas and materials presented must be related to what children already know; in order to acquire mathematical concepts, children need a variety of examples of those concepts in different mathematical forms, different contexts and, possibly, in different modes. Denvir & Brown also stressed the importance of the repetition of the mental processes involved in appropriate tasks in order to develop new mathematical skills.

All these recommendations were drawn upon in the Primary National Strategy's (DfES, 2005) guidance for interventions for children with significant difficulties in mathematics.

In addition to specific teaching strategies, Dowker (2004) recommended ways of assessing children during lessons. Ginsberg (1977) explained that children's mistakes are seldom random, and usually systematically wrong, and that an individual pupil's difficulty may take a unique form. Ginsberg therefore suggested the use of informal methods to gain insight into children's understanding and how we can help them. If we wished to examine a pupil's mathematical understanding, we might ask them to explain their reasoning in order to probe their understanding (Davis, 1984). However, Hiebert & Carpenter (1992) highlighted the difficulty of examining pupils' understanding, and suggested that a variety of opportunities for children to demonstrate their understanding of a mathematical concept was required. Based on the literature, therefore, the desirable characteristics that we looked for in the NC lessons were the following:

Structural characteristics: a focus on specific mathematical difficulties; building on existing knowledge of children and consolidation of earlier learning; practise of fact retrieval; opportunities to solve word problems and practical problems.

Pedagogical characteristics: use of questioning for assessing understanding; use of concrete and visual materials; practical engagement of children in activities; opportunities for discussion, explanation from children and reflection on methods from children; opportunities to develop correct mathematical language.

The characteristics were classed as structural, in terms of incorporating them in the structure of the lesson, and pedagogical in terms of how the NCT interacts with the pupil. We used these characteristics to examine the format of the NC lesson. The recommended structure of the lessons is provided by Edge Hill University and takes the following format (with the descriptions and times as given in the NC handbook for teachers):

Table 8-1: NC lesson description

Learning focus	Minutes (approx)	Description
Familiar activity	4	The child is given a choice from favoured activities, designed to be fun, to set a positive tone to the start of the lesson, and to reinforce existing skills and strengths.
Counting activity	4	The child should have regular opportunities to practise and reinforce a wide range of counting activities, strategies and skills.
Current learning activity 1	8	These two parts of the lesson are where the main teaching takes place. The objectives will be directly linked to the child's individual teaching programme identified through the diagnostic assessment and will be taken from the Weekly Record. This will include key vocabulary and language structures that the child will be supported to use and key questions that the teacher will use to extend learning.
Current learning activity 2	8	The objective for both of these activities will vary depending on the stage of learning, for example Activity 1 might be reinforcing or practicing an objective that the child has progressed in, and Activity 2 would be introducing or extending a new objective. The objectives are specified to this lesson, not more general objectives for the week.
Further applications opportunities	4	Although all parts of the lesson should include 'real-life' application, this section is designed to ensure there is plenty of opportunity to link the learning to real-life scenarios, to exploit different situations, and to illustrate concepts through a variety of situations.
Reflection	2	A dialogue with the child returning to the learning objectives, reflecting on the success criteria and discussing what the next steps should be.

All the lessons observed contained a specific focus, building on previous knowledge and the practise of fact retrieval. Despite the further applications opportunities part of the lesson, only about a quarter of the lessons observed contained opportunities to solve problems.

The Numbers Count handbook outlines key elements of learning and teaching as containing the following: a positive, lively and fun environment; use of a wide range of resources; use of a range of models and images; the teacher making continuous and informed decisions.

We related these to the desired pedagogical characteristics of questioning, use of concrete and visual materials, and practical engagement of children. All the lessons observed contained these characteristics. In addition, all the lessons provided opportunities for children to develop their language.

One desirable characteristic of the lessons, namely the incorporation of opportunities to solve problems, was covered less well than the other characteristics. From the first online survey of NCTs, it was noted that incorporating using and applying tasks in each lesson was one of the areas in which teachers were least confident. Therefore, one possible reason

why this characteristic was not fully covered in NC lessons may have been due to teacher confidence. Also, it was frequently observed that the further applications part of the lesson was not covered, simply because the teacher had run out of time. The other area in which teachers expressed a lack of confidence in the online survey was overall reflection on learning and identification of next steps in each lesson; was sometimes omitted due to lack of time. NCTs were unsure how to cover these aspects, in particular the f part of the lesson, given the constraints of the 30 minutes lesson. Therefore, we can interpret teachers' lack of confidence as uncertainty about covering these aspects, given the constraints of the timings for the lesson. In the second online survey to teachers, a majority of NCTs (54%, n = 142) agreed that the part of the lesson could be removed and incorporated in other parts of the lesson. One way of developing the NC lessons could be to incorporate and throughout the lesson, thereby reducing the time constraints.

8.3.2: Impact on training

Teachers perceived the professional development in a very positive way. From the first online survey of teachers, the most positive aspects of the NC professional development were judged to be opportunities to discuss practice with colleagues (95%, n = 230 found this very valuable); support and visits from TLs (87%, n = 211); and focussing on teachers' understanding of the early mathematics curriculum (84%, n = 203). Opportunity for discussion during the training sessions was considered to be the most valuable part of the professional development. This discussion helped teachers to share ideas for their NC lessons and problems that they were experiencing. The discussion also facilitated learning in the training sessions, and developed the opportunity to reflect on their own practice.

A possible lesson to be learnt from the training concerned the use of video clips. Teachers were asked to record examples of their own practice in NC Lessons and use these as a basis for discussion and reflection during the training sessions. Teachers mentioned both positive and problematic aspects of using the video. The opportunity to examine each others' practice through video was highlighted in the analysis of the first online survey to teachers as being one of the less useful aspects of the professional development (44%, n = 106 finding this *very valuable*). Issues relating to time and technicalities in using the video were highlighted as problematic in the interviews and observations of the training sessions. In some of the training sessions all teachers were asked to bring in a video but there was only sufficient time to watch a few videos. For some teachers there was an overall uncertainty about how to use the videos for their professional development. Teacher leaders also recognised the possible drawbacks to the way that the video had been used in the training. Due to time issues, technical difficulties and uncertainties in using the video, we recommend that support is provided to teachers in using the video for their professional development. Specific technical help could be provided to teachers to enhance the video for the use of reflection as well as more specific guidance on how to use the video.

8.3.3: Wider school level impact

Both parents and head teachers pointed to the many perceived valuable benefits to learning beyond mathematics. These included both specific techniques and strategies which were applicable to other lessons, as well as positive attitudes to education and school generally. For some children simply learning to play in a more positive and constructive way was an

important outcome and one that could well lead to better learning outcomes in other subjects. NC was believed to have helped the children to become more confident and capable learners, and break the cycle of failure associated with learning. Teachers and other staff reported seeing the children in a new light as a result of NC, and there was an expectation that this would have a positive long term impact.

A few of the parents and head teachers pointed to the social difficulties their children experienced and there was a strong feeling that NC helped them to develop and improve their social skills. Many pointed to the NCT as being a positive role model, having the time to show their child how to behave, and stated that this would not happen during normal lessons. It is worth noting that many of the parents we spoke to generally did not view NC as simply an extension of special needs but rather as a programme to get their children back on track. Parents were generally very positive about the changes they saw in their children at home, which they attributed in a large part to NC. Examples included bringing work home and wanting to do it with the parents/carers, and being able to play with their siblings. Improved attendance was mentioned, as was simply being keener in the morning to leave for school. Several parents explained they thought NC had helped improve their own mathematical skills; this included having to help their child with their homework and to do the NC activities, as well as their child showing them how to do certain mathematical techniques. Beyond this, many of the parents spoke very positively about activities some of the schools had organised, including parent workshops, and again how this had helped them begin to overcome their own mathematical difficulties.

Many of the head teachers pointed out that ECC had helped raise the profile of mathematics in their school, and in some instances had provided a useful opportunity to examine mathematics teaching throughout the school in a positive rather than punitive way, for example, by informally showing other staff what was happening and what appeared to be effective, rather than for instance bringing in a consultant who would tend to start with a negative 'what's wrong' approach. Head teachers noted that many staff, teachers and TAs (as well as some governors), had watched NC lessons, and had responded positively and felt they had learnt from the lessons. Indeed a few head teachers said they felt they had learnt a lot themselves about teaching mathematics, in terms of mathematics pedagogy and effective approaches to teaching more generally.

A few head teachers and LAs emphasised that it was important that the NC work should be integrated and co-ordinated with mathematics teaching throughout the school, and was not simply something that was done to a group of children. One head teacher felt the NCT needed to be a strategic lever for change, which she summarised as helping to ensure that current practice was reviewed throughout the school, and ideally not just current practice in mathematics. Finally, a clear message from many head teachers was that they felt that in order to remain effective NC should not be watered down and the resources should be maintained. A few head teachers suggested that ECC and Every Child a Reader (ECaR) could be more closely integrated, and that the possibility of sharing resources and perhaps staffing should be investigated.

8.3.4: Areas for possible development

Long term sustainability

Many of the head teachers and LAs had doubts about the long term viability of the programme, and these concerns increased as the research progressed. Although almost all felt it was an excellent programme that addressed the needs of many of the children, and had significant wider benefits beyond raising mathematics, the majority were not confident about its long term future, in its current form. An important factor for both schools and LAs in the adoption of ECC was the issue that most of the costs were currently met by the government, and without this funding did not think ECC would be viable in their school or LA, although a few schools said they would try and keep doing ECC whatever the funding.

A number of interviewees questioned the value for money of the intervention and several said they would want to know the long term impact (for example key stage 2 (KS2) results) before committing funding. A number of interviewees (head teachers and LA officers) said that if they had the money themselves, they would operate the programme differently.

In reporting below a wide range of responses and ideas, we have worked on the principle that the model remains much the same, and have therefore not included ideas that would substantially change the nature of the programme. As part of our on-going feedback to the programme managers we discussed in an open and positive way these points and are aware that they are actively considering many of the issues. The issues below represent the most significant points raised by the respondents; we suggest these are areas which would benefit from further research supported by pilot studies and small scale exploratory trials.

Flexible delivery

A number of interviewees were interested in a more flexible model of delivery, for example, not necessarily being tied to the termly cycle as some children needed more time and others needed less. It was suggested by a number of respondents that some of the contact time could be allocated to follow up sessions, for example, in a subsequent term. We were told that, for quite a few of the children, follow-up and further support was essential if they were to maintain their gains, and that, ideally, this should be linked to ECC.

Small group working

This was frequently mentioned in the context of challenges and possible developments. A number of interviewees said that, for them, this was the obvious way forward as they had doubts that any government would continue to fund ECC in the current one-to-one format. Several also suggested a hybrid approach with the diagnostics being done individually and following on from that children matched for the teaching phase. There was, however, strong support for the notion that simply doubling up the sessions was not the answer, but rather this essentially different intervention had to be designed from the bottom up.

Target group

There were some concerns that it was difficult in practice to identify the 5% lowest performing children and they suggested more objective methods for selection, for example a

specific test with a specified cut off point. Conversely, it was also suggested that a more flexible and less precise definition should be used, so that the children in most need of support (as defined by the school) would be eligible for the programme. It was also suggested that screening of the whole cohort should be carried out for both numeracy and literacy as there were often common factors between both.

Joined with other initiatives

Several people interviewed suggested that ECC and ECaR could in effect be merged, with teachers trained to deliver both. 4% (n = 10) of the teachers who responded to the questionnaires had also been trained as an ECaR teacher. This would have the advantage of addressing the problem of schools not knowing whether to opt for ECC or ECaR, and it was also suggested that there are many cross-over benefits, for example, implementing one may also help outcomes in the other. This point was taken further by a group of interviewees who suggested that what was needed was a really good specialist catch up teacher supported by well trained TAs for all areas of key stage 1 (KS1) in every school.

Staffing

The issue of contracts, including the 0.5 / 0.6 issue, was raised and concerns about the long term viability of the model were expressed. It was suggested that schools might have difficulty retaining good NCTs and that a clearer career path was needed. Several head teachers also commented on the amount of time their NCTs were out of school and suggested that more training and meetings should be done outside of the school day or in the school holidays.

Para-professionals

There were a number of suggestions as to how TAs and other para-professional staff could be used more effectively, including for screening children and exit testing. There was also a suggestion that TAs could be trained to a similar level as NCTs and deliver the programme themselves. One school was looking at a teacher led, TA delivered model, and several schools had TAs working alongside NCTs so that they could take over should the funding be cut. Several head teachers suggested that more formal use could be made of TAs for the re-integration and further classroom support of children.

Peripatetic model

Several LAs advocated the development of a peripatetic model of delivery, with a number of NCTs employed centrally, going out to schools as required. Whilst a number of head teachers also explored this notion, and broadly accepted the logic, the main argument against it was that many of the wider/whole school benefits might be lost without the appointment and continuity of one a specific NCT.

8.4: Conclusions

8.4.1: Key features of effective implementation of Numbers Count

A number of research studies on mathematics interventions (Dowker, 2004; Denvir & Brown, 1986) have previously identified a number of desirable structural characteristics of lessons. These include focusing on specific mathematical difficulties, building on existing knowledge and consolidation of earlier learning, practise of fact retrieval, and opportunities to solve word problems and practical problems. In addition, identified desirable pedagogical characteristics include the use of questioning for assessing understanding, use of concrete and visual materials, practical engagement of children in activities, opportunities for discussion, explanation from children and reflection on methods from children, and opportunities to develop correct mathematical language.

Virtually all of the NC lessons we observed contained these characteristics, although only about a quarter contained opportunities to solve problems. Feedback from NCTs highlighted the difficulties in covering aspects such as the further applications part of the lesson given the time constraint of a 30 minute lesson.

NCTs perceived their professional development in a very positive way, the most valuable aspect of the NC professional development being the discussion that took place between colleagues during these sessions and facilitated by the experience that the trainer brought to the sessions. From the first survey and interviews with NCTs, the support that the TLs and NTs provided, and the value of the visits that NCTs received from TLs was emphasised as particularly positive aspects of the professional development.

Head teachers and LA officers stressed the importance of NCTs being well trained and well prepared, and almost without exception both groups were highly positive about the professional development for NCTs. Head teachers generally were happy that they knew what would be required of them and their schools to support NC, although there were calls for more on-going support meetings at a cluster or LA level, typically as part of existing routine meetings. Concerns were expressed by head teachers about potential recruitment difficulties. Several schools had experienced problems locally, and it was suggested that this might be worse if the programme were to be rolled out nationally. It was felt that the working arrangements such as short term contracts and lack of a career progression could be a disincentive to good candidates.

NCTs needed a degree of support from school managers, but also a degree of freedom. NC delivery was perceived to require a certain amount of entrepreneurial initiative, for example, finding ways of engaging challenging children and their families, and being able to work positively and creatively with colleagues. To this end, head teachers often suggested that a flexible and collegiate style of management was best suited, as well as one that promoted inclusion of the NCT in the overall school management and organisation process, if schools were to benefit from what was recognised as highly skilled teachers.

8.4.2: Key features of the effective implementation of a small group intervention model.

Perhaps the most important point made in the course of our investigation was that pairs work is not simply double ones but rather a somewhat different intervention, albeit with common aims and objectives. It should also be noted that we could only observe a relatively small number of small group lessons (mainly pairs), and this, combined with an inevitable possible selection bias (schools were free to choose to participate in the pairs trial), makes it more difficult to generalise from the findings. Therefore the findings should be viewed as indicative, rather than as definitive.

From our observations the advantages centred on two key areas: cooperative learning, and preparation for re-integration. We stress that many of the features were present in lessons with individual children, in that pairs working was quite often a feature towards the end of the programme. In practice, many of the techniques and benefits of peer learning or mentoring were evident.

We also observed a number of potential disadvantages of pairs working, although depending on the nature of the children and the pairings. It should be noted that the pairings were not necessarily ideal, being constrained by the need to randomize the allocations to pairs for the trial (although as mentioned earlier the process had two steps and the teachers made the decisions in the second step, see Trial 2: Pairs in Chapter 3). Had schools been entirely free to choose the pairs this may have affected the responses. The disadvantages centred on issues such as increased need to differentiate during the lessons, and some pupils dominating the lessons and not being able to focus very tightly on a particular issue for a specific child.

At the organisational level there were many benefits of pairs and small group working, as well as some notes of caution. The most common response was that such working needed to be approached with a degree of flexibility. For example, most head teachers and LA officers saw the value of carrying out the diagnostic process on a one-to-one basis, indeed some went further and said that this was the only way it should be done. We were informed that the groupings were critical (even if several schools were pleasantly surprised as to how unlikely pairings worked out), and that schools should support the NCT to make alterations to the groupings as appropriate.

Parents[5] voiced concerns that the important and valuable relationship that their child had formed with the NCT might be put at risk from unsuitable pairings. That said, some recognised the possible advantages of their children learning to work with other children, which in some cases had previously been difficult for them.

Other advantages of pairs and small group working were also cited. Firstly, the cost advantage was discussed by many head teachers and LAs, although they thought this had to be carefully weighed against the potential downsides which might lead to less progress overall. As we had observed in the lessons, there were also positive comments about the advantages to learning of working in pairs (or small groups), for example, many were familiar with peer tutoring and could see the relevance (and advantages) of such an approach.

[5] We only spoke with parents from children receiving individual NC, and not any working in pairs.

8.4.3: Key factors that enable teachers trained to delivery ECC to have a wider impact on learning, teaching and mathematics standards in their schools

In this section we focus on the perspective of the NCTs, and it should be noted that these may not necessarily appear to coincide with the responses from the head teachers and parents. Year 2 teachers and teaching assistants (TAs) highlighted the benefits that they had gained from liaising with the NCTs, for example benefits with regard to tracking pupils' progress, or the NCTs being a source of ideas and resources for teaching mathematics. NCTs themselves felt that their role was beneficial to other staff as they had been able to share the expertise that they had developed from the NC training and their experience of dealing with specific mathematics difficulties.

With regard to barriers that existed for their wider role, the first survey of NCTs found that almost a quarter of them had had no formal opportunities to liaise with colleagues in school. The issue of time was important for the NCTs in terms of liaising with other staff. Year 2 teachers and TAs also identified the need for more time for discussions with the NCT. Many NCTs said they thought that a 0.5 or 0.6 position was insufficient for the broader role, and the issue of combining the NC role with another role was a frequently raised difficulty. In order to support the progress of pupils after the NC intervention, the second survey of NCTs found that teachers felt that the most valuable suggestions would be more liaison with and information and training for TAs and Year 2 teachers.

Teachers thought that parental involvement and its link with pupil progress in the intervention was very important. Year 2 teachers and TAs also highlighted the possible benefits of the programme on parents, in particular in terms of confidence in working on mathematics with their children. However, the first survey to NCTs found that the vast majority of respondents (69%) thought that less than a quarter of parents of NC pupils had observed lessons; however, parents were thought to be more involved in mathematical activities which were sent home with the children. Barriers for involving parents included work commitments of the parents, the existence of younger siblings to care for, language difficulties including speaking a language other than English, and a general feeling of intimidation including an awareness of their own difficulties with mathematics.

8.4.4: Challenges to the effective implementation of the programme

The overwhelming challenge identified by head teachers and LA officers was sustainability. The messages about the long term funding and future of ECC towards the end of the research period were not very positive, and many were contemplating various options. Whilst there were signs of possible targeted funding very few felt they could continue with ECC using only their core funding, and likewise LAs were not confident about alternative funding streams. Whilst this may have influenced people's perceptions and responses, it may have also helped focus minds and through this lens we summarise below some of the key and recurring responses presented as areas for further research and consideration, and not as complete solutions.

The key issue was perceived as being able to target the intervention at the most appropriate children, namely the bottom 5% nationally. Many respondents pointed out the difficulties of identifying this population, and whilst there was heavy reliance on teachers' perceptions, it

was acknowledged that these could be inaccurate, although this was very rarely reported. This was due, for example, to children developing at different rates, and the limitations of subjective assessments. Objective screening tests do not provide a simple answer either, as they too are susceptible to identification false positives (and negatives). The general consensus was that a combination of screening and teacher perceptions was the most appropriate way to identify the target 5%, but the challenge of how to provide ECC to all of these children (approximately 30,000) remains. The Williams review (2008) addresses this point in some detail. We were looking at the development phase of ECC, but for a few respondents there was some doubt that the current model, if rolled out, would achieve its aims. The biggest problem was the unequal distribution of this bottom 5%. Some schools might have 18 eligible pupils, others 12 and some just one or two. A variety of suggestions was offered concerning how this issue might be addressed.

The key feature underpinning most of the responses was for greater flexibility of delivery. This would not have been appropriate during the development phase, beyond the pairs/small group working, as it would have had a confounding effect on the findings of the trials. As already mentioned, there was strong support, in principle, for pairs/small group working, whilst maintaining one-to-one working for the diagnostic element, as well as an option to provide one-to-ne NC in specific cases where needed. It was also suggested that ECC need not necessarily be offered only in a one term block, and that children could start at different times of the term and stay in the programme for more variable amounts of time, albeit with safeguards to ensure the fidelity of the programme remained intact.

The use of different staffing models was discussed extensively. For example, some schools and LAs felt that TAs, in particular higher level TAs, could be trained to deliver ECC, possibly under the supervision of the mathematics co-ordinator. There was also support for a peripatetic model. This approach was not thought to be without its downsides, though. For example, many of the whole school/wider benefits might be lost. Finally, a number of respondents suggested that suitably trained teachers could deliver both ECC and ECaR. Some teachers are already qualified to deliver both. It was proposed that this would help address both the small numbers issue, as well as what was perceived by head teachers from small and medium-sized schools to be an impossible choice, namely should they opt for ECaR or ECC?

Chapter 9: Conclusions

The results from Trial 1 demonstrated that Numbers Count (NC) had a moderate short-term impact on mathematical abilities. The results from Trial 2: Pairs found that teaching adapted NC in pairs may have a similar level of impact. However, we were unable to confirm the short- or medium-term impact of ECC when looking at the key stage 1 results in the secondary analyses. The process evaluation found ECC to be a well designed, highly regarded and effectively implemented programme. The costs of the delivering the programme (one-to one) are relatively high compared to other mathematics interventions. In essence, whilst we found that NC is able to improve children's mathematical skills, the relative cost may preclude it as a realistic option for many schools.

Our key recommendation, therefore, is that future research should support the identification and development of a range of interventions aimed at reducing underachievement in mathematics, and that robust evaluations should provide reliable and high quality evidence of their effectiveness. In practice this could be achieved by a series of randomized controlled trials to assess impact, accompanied by economic evaluations to assess the cost of achieving the given level of impact, and process evaluations to consider their implementation.

References

Altman, D.G., Schulz, K.F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., Gotzsche, P.C. and Lang, T. (2001) 'The revised CONSORT statement for reporting randomized trials: Explanation and elaboration' *Annals of Internal Medicine* 134 (8): 663-694.

Blatchford, P., Galton, M., & Kutnick, P. (2005) *Improving the effectiveness of pupil groupings in classrooms* ESRC/TLRP final project report. Swindon: ESRC.

Clause-May, T., Vappula, H. and Ruddock, G. (2004) *Progress in Mathematics 6* GL Assessment.

Davis, R.B. (1984) *Learning Mathematics: The Cognitive Approach to Mathematics Education*. London: Croom Helm.

Denvir, B., & Brown, M. (1986) Understanding of number concepts in low attaining 7-9 year olds: Part I. Development of descriptive framework and diagnostic instrument. *Educational Studies in Mathematics*, 17(1), 15-36.

DfES (2005) *Targeting support: implementing interventions for children with significant difficulties in mathematics*. DfES Publications.

DfES (2006) *Primary framework for literacy and mathematics – Guidance paper – Using and applying mathematics*. DfES Publications.

Douetil, J. (2004) *The long term effects of Reading Recovery on National Curriculum tests at end of Key stages 1 and 2.* London: Institute of Education.

Dowker, A. (2004) *What works for children with mathematical difficulties?* Nottingham: DfES publications.

Drummond, M. F., Sculpher, M. J., Torrance, G. W., O'Brien, B. & Stoddart, G. L. (2005) *Methods for the Economic Evaluation of Health Care Programmes.* Oxford: Oxford University Press.

Edge Hill University, Lancashire County Council and Every Child Counts (2008) *Numbers Count Handbook 2008 – 2009* Edge Hill University.

Every Child a Chance Trust (2009) *Long term costs of literacy difficulties 2nd Edition* London, Every Child a Chance Trust.
Accessed at:
http://www.everychildachancetrust.org/ecar/pubs/long_term_costs_of_literacy_report.pdf

Every Child a Chance Trust (2009) *Long term costs of numeracy difficulties* London, Every Child a Chance Trust.
Accessed at:
http://www.everychildachancetrust.org/pubs/ECC_long_term_costs_numeracy_difficulties_final.pdf

Ginsberg, H. (1977) 'Learning difficulties.' In A. Floyd (ed.), *Developing mathematical thinking* (pp. 162-176). London: Open University Press.

Goodman, R. (2005) *Strengths and Difficulties Questionnaire* Youth in Mind Accessed and downloaded from: http://www.SDqinfo.com/

Haylock, D. (1991) *Teaching mathematics to low attainers, 8-12*. London: Paul Chapman Publishing Ltd.

Hiebert, J., & Carpenter, T.P. (1992) 'Learning and teaching with understanding.' In D. A. Grouws (ed.), *Handbook of research on mathematics teaching and learning* (pp. 65-97). New York: Macmillan.

Lipsey, M.W. and Wilson, D.B. (1993) 'The efficacy of psychological, educational and behavioral treatment: Confirmation from meta-analysis' *American Psychologist* 48 (12): 1181-1209.

Moher, D., Schulz, K. F. and Altman, D. G. (2001) 'The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials' *BMC Medical Research Methodology* 1:2.

Slavin, R and Madden, N. (2008) *Understanding Bias Due to Measures Inherent to Treatments in Systematic Reviews in Education* Paper presented at the annual meetings of the Society for Research on Effective Education, Crystal City, Virginia, March 3-2008.

Torgerson, C., Wiggins, A., Torgerson, D.J., Ainsworth, H., Barmby, P., Hewitt, C., Jones, K., Hendry, V., Askew, M., Bland, M., Coe, R., Higgins, S., Hodgen, J., Hulme, C., Tymms, P. (2011a) Every Child Counts: The independent evaluation, Executive summary, report to DfE, January 2011.

Torgerson, C., Wiggins, A., Torgerson, D.J., Ainsworth, H., Barmby, P., Hewitt, C., Jones, K., Hendry, V., Askew, M., Bland, M., Coe, R., Higgins, S., Hodgen, J., Hulme, C., Tymms, P. (2011c) Every Child Counts: The independent evaluation, Appendices, report to DfE, January 2011.

Williams, P. (2008) *Independent Review of Mathematics Teaching in Early Years Settings and Primary Schools: Independent Review Final Review* June 2008 DfES London Crown Copyright.

Ref: DFE-RR091a

ISBN: 978-1-84775-873-6

March 2011