The effectiveness of an intensive individual tutoring programme (*Numbers Count*) delivered individually or to small groups of children: A randomised controlled trial

Torgerson, C., Wiggins, A., Torgerson, D., Ainsworth, H., Hewitt, C on behalf of the *Every Child Counts* independent evaluation team*

Author contact details
Corresponding author: Professor Carole Torgerson: carole.torgerson@durham.ac.uk Durham University, School of Education, Leazes Road, Durham, DH1 1TA

Dr Andy Wiggins: andy.wiggins@durham.ac.uk Durham University, CEM, Mountjoy Research Centre 4, Stockton Road, Durham, DH1 3UZ
Professor David Torgerson: david.torgerson@york.ac.uk University of York, Dept. of Health Sciences, Heslington, York, YO10 5DD
Mrs Hannah Ainsworth: hannah.ainsworth@york.ac.uk University of York, Dept. of Health Sciences, Heslington, York, YO10 5DD
Dr Catherine Hewitt: catherine.hewitt@york.ac.uk University of York, Dept. of Health Sciences, Heslington, York, YO10 5DD

Word count: 4992

*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. And Tymms, P.

Abstract

Introduction
In this paper we present the results of two small randomised controlled trials (RCTs) investigating the relative efficacy of a one-to-one numeracy programme, *Numbers Count* (NC). This intervention was developed as part of the *Every Child Counts* (ECC) programme to specifically target the lowest achieving children at key stage 1 (KS1) mathematics. The research question focused on the relative efficacy of NC delivered either individually (as originally intended by the developers of the programme) or in an adapted version to small groups of two or three children.
Design and methods
In 15 schools 75 children in year 2 identified by the schools as being eligible to receive *Numbers Count* were randomized to receive it individually or in pairs and in 7 schools 54 children in year 2 identified by the schools as being eligible to receive *Numbers Count* were randomized to receive it individually or in triplets during the school year 2009-10. The design of the trial required 5 children, for the pairs sub-trial, or 7 children for the triplets sub-trial *in each school* to receive NC individually or in pairs/triplets in autumn term 2009 or spring term 2010.
Results
The primary outcome measure Progress in Maths 6 (PIM 6) was undertaken and marked blind to group allocation by independent testers. We found 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. We also found no statistically significant differences between children taught individually or in triplets in terms of PIM 6 scores. We pooled the effect sizes for individual versus pairs and triplets delivery in a meta-analysis of individual versus small group teaching. The pooled effect size was -0.26 (CI -2.18 to 1.65) which demonstrates no statistically significant difference between individual and small group teaching.
Conclusion
There was no evidence of a difference between the groups, which was not unexpected as our sample sizes were relatively small. This is the only robust evidence from RCTs of the promise of the intervention. However, given the small sample sizes of the trials, we recommend further larger trials should be undertaken comparing small group mathematics teaching with one to one educational interventions, using both this programme and others These two trials could also usefully be included in meta-analyses comparing small group and one-to-one teaching in mathematics education.

Keywords: experimental design; *Numbers Count*; mathematics education; education research

Introduction

The Williams review (2008) of mathematics teaching in early years settings and primary schools, reported that, since 1998 the number of children in the United Kingdom failing to achieve level 3 in mathematics at key stage 1 (KS 1) had remained constant at about 6%. During this same period, the performance of the majority of children had improved. National concern at this situation prompted the then Labour government together with a consortium of charitable trusts (The Every Child a Chance Trust) to establish the *Every Child Counts* (ECC) programme. The aim of this consortium was to address the mathematical needs of the lowest achieving 6% of children at KS 1.

In this paper we present the results of two small randomised controlled trials (RCTs) investigating the relative efficacy of a one-to-one numeracy programme, *Numbers Count* (NC). This intervention was developed as part of the ECC programme to specifically target the lowest achieving children at KS 1 mathematics. The research question focused on the relative efficacy of NC delivered either individually (as originally intended by the developers of the programme) or in an adapted version to small groups of two or three children. It was a recommendation of the Williams (2008) review that research should be conducted to establish whether individual or small group delivery of an intensive numeracy intervention is more effective. The trials were 'pragmatic' (Schwartz and Lelouch, 1967) in the sense that the research question was of practical significance because the intervention was expensive and was in widespread use at the time, and yet no rigorous evidence of its efficacy using an experimental design had yet been conducted.

Independent evaluation of Every Child Counts

In the school year 2009-10 we undertook an independent evaluation of the overarching programme *Every Child Counts.* Much of the underlying pedagogic rationale of ECC was informed by a Department for Education-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 underperform in mathematics are highly susceptible to targeted intervention, and that this should take place at an early age to reduce negative attitudes and allow access to other aspects of the curriculum. *Numbers Count* (NC) is the intensive intervention element of ECC. Its development was led by Edge hill University, which provided the professional development of the specially trained teachers employed by the schools to deliver the intervention.

Numbers Count

The intervention is a 12-week programme which consists of daily 30 minute one-to-one sessions for the target children, delivered by Numbers Count teachers (NCTs). The core elements of the programme are: an initial comprehensive diagnostic assessment of each child's strengths and weaknesses; core learning objectives for the NC lessons; and individual guidance for teachers on lesson structure and key teaching approaches. NCTs are supported by a continuing professional development programme and a quality assurance system (provided by Edge hill University). NC is specifically designed to help children to develop their knowledge and understanding of number, and 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 normal class mathematics lessons. They use shape, space and measures, and handling data as contexts for the development and application of children's number skills. Alongside the intervention, the children continue to study the full breadth of the mathematics curriculum with their class teacher.

Adapted Numbers Count
The NCTs received a small amount of extra training for the small group work and taught fewer sessions (but had more children overall). The make-up of the small groups (pairs or triplets) was determined through a combination of randomization and teacher decision. The teachers involved in delivering NC to children in small groups received minimal professional development on how to adapt NC to enable it to be delivered to two or three children at the same time. The professional development was undertaken by the ECC Trust. It involved training in pupil interaction: pupils listening to others' answers and explanations; learning to take turns and to work with others; 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. The small group lessons tended to be longer than standard NC lessons (45 minutes).

The independent evaluation included an RCT comparing NC with normal classroom teaching. In this first Trial (referred to as Trial 1 and reported to the funder - DfE - Torgerson et al., 2011), we evaluated the effectiveness of receiving NC in addition to normal classroom practice compared with normal classroom practice alone for children's attainment in mathematics using a pragmatic design. In this design 12 children within each of 44 schools were randomly allocated to be in the *intervention group* and receive NC in the autumn term in addition to normal classroom practice or to be in the *control group* and receive normal classroom teaching only in the autumn term and be placed on a NC waiting list to receive NC in the spring or summer terms. In our second Trial (referred to as Trial 2), reported in detail in this paper, our primary aim was to obtain robust evidence of the *relative effectiveness* of *Numbers Count* (NC) delivered *individually to one child* or to *groups of pairs or triplets of children* on attainment in mathematics. We undertook a pragmatic randomized controlled trial (RCT) in 22 schools: 15 schools in the pairs sub-trial and 7 schools in the triplets sub-trial. The trial was funded in 2009 by the then-Department for Children, Schools and Families (DCSF) and has previously been reported in detail to the funder – DfE - Torgerson et al. 2011.

Design and methods

The two sub-trials were designed and reported to comply with the CONSORT statement, which is a mandatory guide adopted by most high ranking medical journals for the reporting of medical trials (Schulz et al, 2010), and is easily adapted for the reporting of educational and social science trials (Torgerson and Torgerson, 2008; Boutron et al 2010).

Recruitment
We identified 22 schools which were willing to take part in the trial (15 in pairs sub-trial and 7 in triplets sub-trial). We asked each school if they would identify children in year 2 who would be eligible to receive *Numbers Count* (assessed by their school as struggling in mathematics).

Randomisation
Children in year 2 identified by the schools as being eligible to receive *Numbers Count* were randomized to receive it individually or in pairs or triplets during the school year 2009-10.

Randomisation was conducted using a bespoke computer programme, which allocated children to receive the intervention firstly individually or in pairs or triplets and secondly into term of delivery. The children's identification numbers were entered into the programme, by the trial manager, which then randomized children as a block in a 2:1 ratio (for pairs) or a 3:1 ratio (for triplets). The programme recorded the identification numbers of children allocated and put these into the trial database, which meant the allocation was independent and secure. The importance of having independent secure randomisation has been highlighted on a number of occasions. RCTs that do not use independent allocation tend to produce larger effect sizes compared with those that do (Hewitt et al, 2009; Hewitt et al, 2005; Schulz et al, 1995). Indeed, the phenomenon of small trials producing larger effect sizes than large trials can be explained by the fact that small trials more often use inadequate allocation methods (Kjaergaard et al, 2001).

Design

The design of the *trial required 5 children, for the pairs sub-trial, or 7 children for the triplets sub-trial* in each school to receive NC individually or in pairs/triplets in autumn term 2009 or spring term 2010, [and 2, 3 or 4 children to receive NC in the summer term 2010 (but note these children were not included in the analysis)]. In the autumn term the Numbers Count teacher delivered NC to one child individually and to 4 (pairs) or 6 (triplets) children in two groups. In the spring term the NC teacher delivered NC to one child individually and to 4 children in two pairs or 6 (triplets) children in two groups. [In the summer term NC was delivered to 2, 3 or 4 children individually (but note these children were not included in the analysis).]

Sample size and power

Our sample size for the pairs sub-trial was driven by the number of schools that were recruited; consequently, we discuss below the likely power we would have achieved from the expected numbers we anticipated we would recruit. In the statistical 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. 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. Note this calculation is illustrative and serves to show that this trial is not adequately powered to find the smaller difference that we would anticipate could exist between the two treatment groups. Similarly for the triplets sub-trial we could only recruit 8 schools with 128 children, which gave us limited power to detect small but educationally worthwhile differences.

Outcome measurement

The pre-test was the Sandwell (A) test; the primary outcome measure was GL Assessment Progress In Maths 6 (PIM 6) which was administered to all children (and marked independently) in January 2010 and in April 2010.

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 (Torgerson and Torgerson, 2003). 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 descriptively. No formal statistical comparisons were undertaken for the baseline data. As above, the primary outcome was the PIM 6: the scores on the PIM 6 were summarised descriptively (means and standard deviations) 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.

The primary analysis compared the children receiving NC individually or in pairs/triplets in the autumn and spring terms (January and April testing respectively), and the results were combined. In total 23 schools consented to take part in the trial and children within 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. 22 schools remained in the trial in April 2010. In the pairs sub-trial a total of 190 children selected by schools with parental/carer consent participated at the beginning of the trial. At the end of the autumn term 189 children remained in the trial, and at the end of the spring term 175 children remained in the trial. In the triplets sub-trial 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 Bank' but note from this point on were not considered as part of the trial, leaving 129 children in the trial.] At the end of the autumn term 127 children remained in the trial, and at the end of the spring term 124 children remained in the trial.

Results

Table 1 gives the baseline characteristics of the children included in the pairs and triplets sub-trials 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 individual or pair. [Four children were randomized to receive NC in the spring or summer term only and have been excluded from the summaries below.]

INSERT TABLE 1 ABOUT HERE

We can see from Table 1 that the groups formed at baseline were comparable in age, Sandwell A mathematics scores, percentage of children receiving free school meals and gender.

Primary outcome
The primary outcome measure (PIM 6) was undertaken and marked blind to group allocation by independent testers.

INSERT TABLE 2 ABOUT HERE

January testing – pairs sub-trial
The mean PIM 6 score for the children receiving the NC intervention individually was 15.5 (SD 3.5) and for the children receiving 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 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 2).

January testing – triplets sub-trial
The mean PIM 6 score for the children receiving NC individually was 17.5 (SD 4.3) and for the children receiving NC in triplets was 18.0 (SD 4.4). The results highlight the finding that children who received 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 2).

April testing pairs sub-trial
The mean PIM 6 score for the children receiving the NC intervention individually was 17.8 (SD 4.8) and for the children receiving 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 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 2).

April testing – triplets sub-trial
The mean PIM 6 score for the children receiving NC individually was 22.8 (SD 3.4) and for the children receiving NC in triplets was 19.2 (SD 5.4). The results highlight the finding that children who received 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 2).

Overall - pairs
In Figure 1 we combine the two analyses of individual and pairs delivery and demonstrate 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.

INSERT FIGURE 1 ABOUT HERE

In Figure 2 we combine the two analyses of individual teaching versus teaching in triplets. As the figure demonstrates, the pooling shows no statistically significant difference between the scores of the children taught individually or in triplets in terms of PIM 6 scores.

INSERT FIGURE 2 ABOUT HERE

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, and therefore the results should be treated with caution.

INSERT TABLE 3 ABOUT HERE

The effect size for the pairs sub-trial in 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). For the triplets sub-trial 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 in either of the two sub-trials.

Meta-analysis

In figure 3 we pooled the effect sizes for individual versus pairs and triplets delivery in a meta-analysis of individual versus small group teaching. The pooled effect size was -0.26 (CI -2.18 to 1.65) which demonstrates no statistically significant difference between individual and small group teaching.

INSERT FIGURE 3 ABOUT HERE

Exploratory analysis

For the primary outcome measure (PIM 6) we also explored whether children responded differently to NC based upon a number of pupil characteristics.

January testing – pairs sub-trial

From the analysis, we found little or no evidence to suggest that children responded differently to NC based upon any of the pre-specified interactions: pre-test scores (p=0.27), age (p=0.26), gender (p=0.91) or free school meals (p=0.09).

January testing – triplets sub-trial

From the analysis, we found little or no evidence to suggest that children responded differently to NC based upon most of the pre-specified interactions: pre-test scores (p=0.01), age (p=0.81), gender (p=0.02), free school meals (p=0.18).

April testing – pairs sub-trial

From the analysis, we found little or no evidence to suggest that children responded differently to NC based upon most of the pre-specified interactions: pre-test scores (p=0.78), age (p=0.95), gender (p=0.04) or free school meals (p=0.40).

April testing – triplets sub-trial

From the analysis, we found little or no evidence to suggest that children responded differently to NC based upon all of the pre-specified interactions: pre-test scores (p=0.67), age (p=0.53), gender (p=0.63), free school meals (p=0.15).

Discussion

In Trial 2 we undertook two sub- trials comparing the effect of individually delivered NC compared with delivery in pairs or triplets. We found no statistically significant difference between the individual or small groups delivery. However, as we noted earlier these trials are underpowered to demonstrate modest but important differences between the modes of teaching. The larger trial of NC reported separately (Trial 1) showed an effect size difference of 0.30. Comparing NC in pairs or triplets to NC as individuals we would expect to see an even smaller difference. Consequently, both of the trials described here were underpowered. Therefore, it is important that larger trials are undertaken to compare individual and small group delivery of the intervention, and as such this would have policy implications for schools considering one to one tuition.

This study has a number of strengths and limitations. The key strengths are robust design conforming to reporting recommendations in the CONSORT statement, robust randomization

procedures for the trials and independent testing (Torgerson and Torgerson, 2003). We present CONSORT flow-diagrams for both sub-trials in Appendix A. However, we had relatively few participants leading to relatively wide confidence intervals around the estimates of effect. In summary, our data, within the limitations of the small sample sizes, suggest that NC delivered in pairs or triplets is similar in effectiveness to individual delivery of NC. Therefore, until there is evidence to the contrary we would recommend, on the basis of costs, that NC should be delivered in small groups of at least two children rather than on a one to one basis.

Interaction analysis

It may be possible, however, for one to one delivery to be more effective among children with certain characteristics. However, we found little or no evidence to suggest that children responded differently to NC based upon most of the pre-specified interactions. The only statistically significant interaction identified was for gender: in the April testing period for the pairs sub-trial and in the January testing period for the triplets sub-trial. However, due to the large number of analyses being undertaken and the inconsistency across terms this finding is likely to be due to chance rather than due to a 'true' effect.

Conclusion

There was no evidence of a difference between the groups, which was not unexpected as our sample sizes were relatively small. There was no evidence of a difference between the groups, which was not unexpected as our sample sizes were relatively small. This is the only robust evidence from RCTs of the promise of the intervention. However, given the small sample sizes of the trials, we recommend further larger trials should be undertaken comparing small group mathematics teaching with one to one educational interventions, using both this programme and others These two trials could also usefully be included in meta-analyses comparing small group and one-to-one teaching in mathematics education.

We were able to use the existing infrastructure of the intervention *(Numbers Count)* to keep the costs of undertaking the two additional sub-trials low. In addition, it was clear that individual delivery of this expensive intervention was unlikely to be financially sustainable in the long-run, and therefore, it was important to explore the feasibility of adapting the intervention to small group delivery and test its differential impact. Given the findings that small group (pair or triplet) delivery of NC was similar in effectiveness to individual delivery, our results could be used in power calculations for a larger trial of individual versus small group teaching.

Acknowledgements

The trial described in this article was funded by the Department for Children, Schools and Families (DCFS). We also acknowledge the University of York and Durham University for additional funding to support the trial. The corresponding author held posts at the Universities of York and Birmingham during the period when the research took place and currently holds a position at Durham University.

References

Torgerson D.J., Torgerson, C.J. (2003) Avoiding bias in randomised controlled trials in educational research, *British Journal of Educational Studies*, 51: 36-45

Torgerson, D. and Torgerson, C. Designing and running randomised trials in health, education and the social sciences, Palgrave Macmillan, March, 2008

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. (2011) The Every Child Counts Independent Evaluation Report to DfE, 31st Mar., 2011

Boutron I, John P, Torgerson DJ. Reporting methodological items in randomized experiments in political science. *Annals of the American Academy of Political and Social Science* 2010;628**:**112-131.

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

Hewitt C, Hahn S, Torgerson DJ, Watson J, Bland JM. (2005) Adequacy and reporting of allocation concealment: review of recent trials published in four general medical journals. *British Medical Journal* 330, 1057-1058.

Hewitt CE, Torgerson DJ, Berger V. Potential for technical errors and subverted allocation can be reduced if certain guidelines are followed: Examples from a web-based survey. *Journal of Clinical Epidemiology* 2009;62;261-269

Kjaergaard LL, Villumsen J, Cluud C. Reported Methodologic (2001) Quality and discrepancies between large and small randomized trials in meta-analyses. *Annals of Internal Medicine* 135, 982-989.

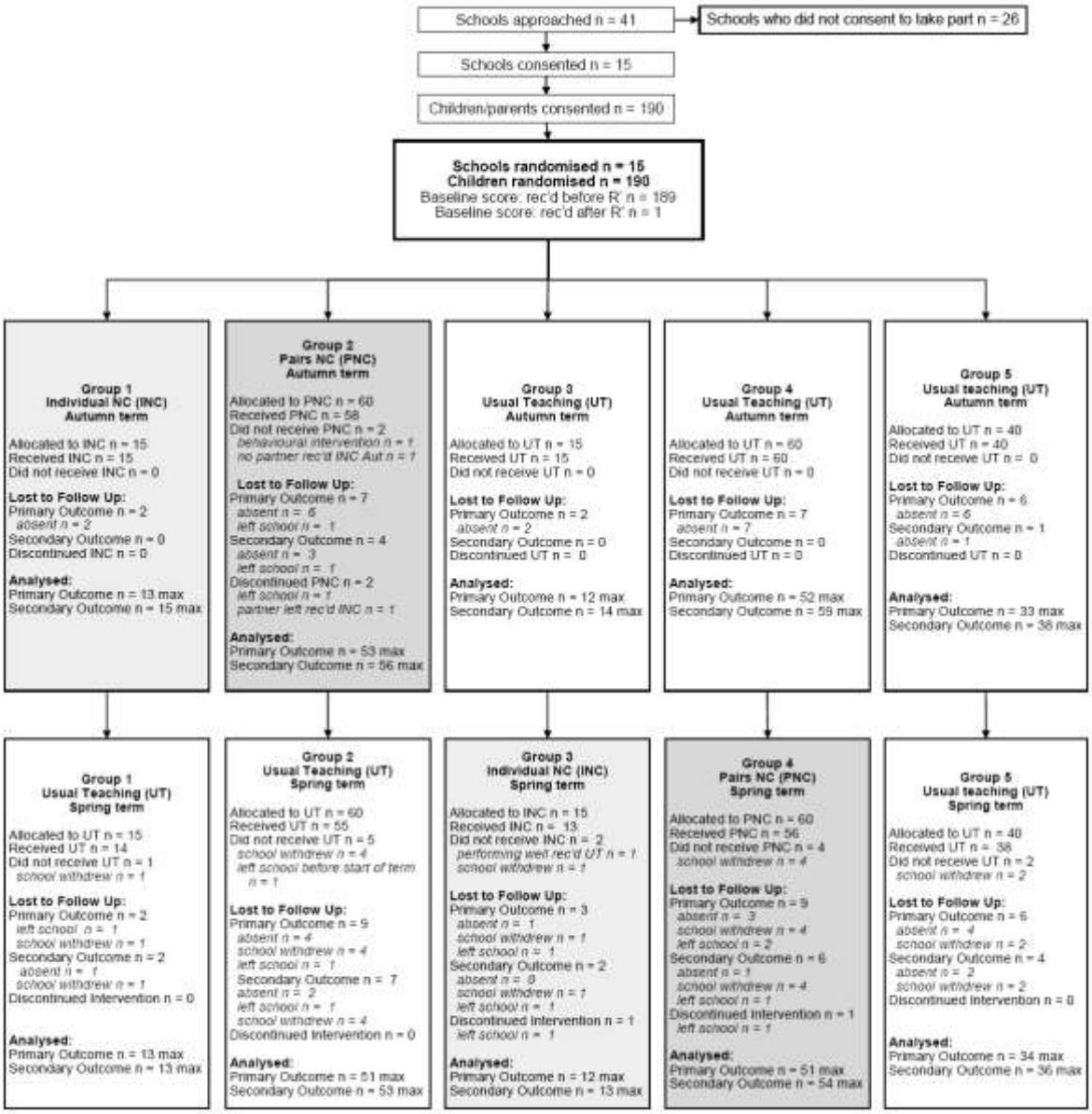
Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010; 340:c332

Schulz KF, Chalmers I, Hayes RJ, Altman DG. (1995) Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of the American Medical Association* 273, 408-412.

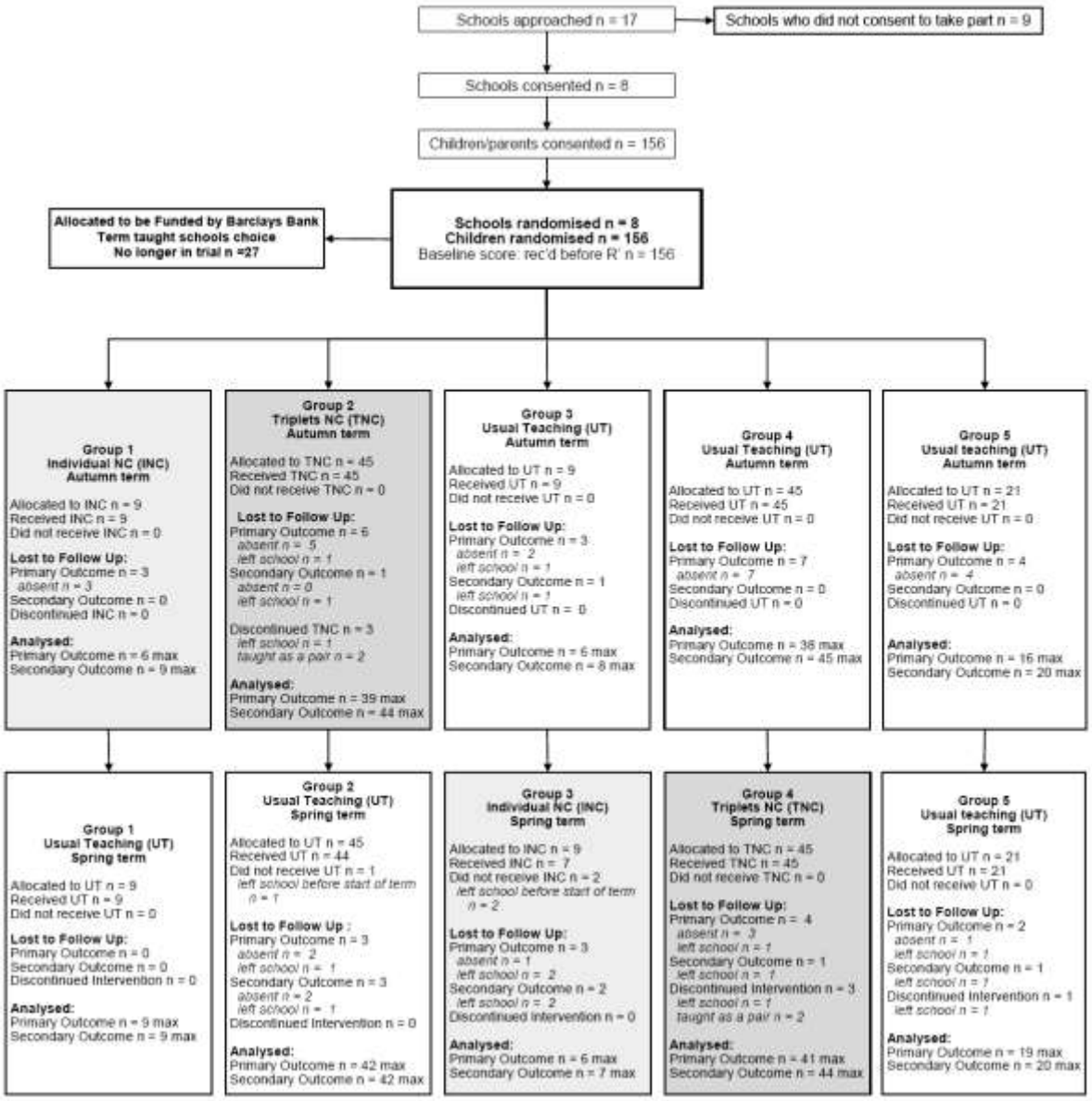
Schwartz, D., Lelouch, J. (1967) Explanatory and pragmatic attitudes in therapeutic trials. Journal of Chronic Disease; 20: 637-648

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

APPENDIX A
Pairs sub-trial CONSORT flow diagram



Triplets sub-trial CONSORT flow diagram



TABLES AND FIGURES

Table 1: Baseline characteristics for pairs and triplets sub-trials

Pairs sub- trial

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)

Triplets sub-trial

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)

Table 2: Primary outcome measure for pairs and triplets sub-trials

Pairs sub-trial				
Outcome	**Individual**	**Pairs**	**Estimate***	**ES**
PIM6 (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)
PIM6 (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)
Triplets sub-trial				
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.

Figure 1 Forest plot of NC individual versus pairs delivery

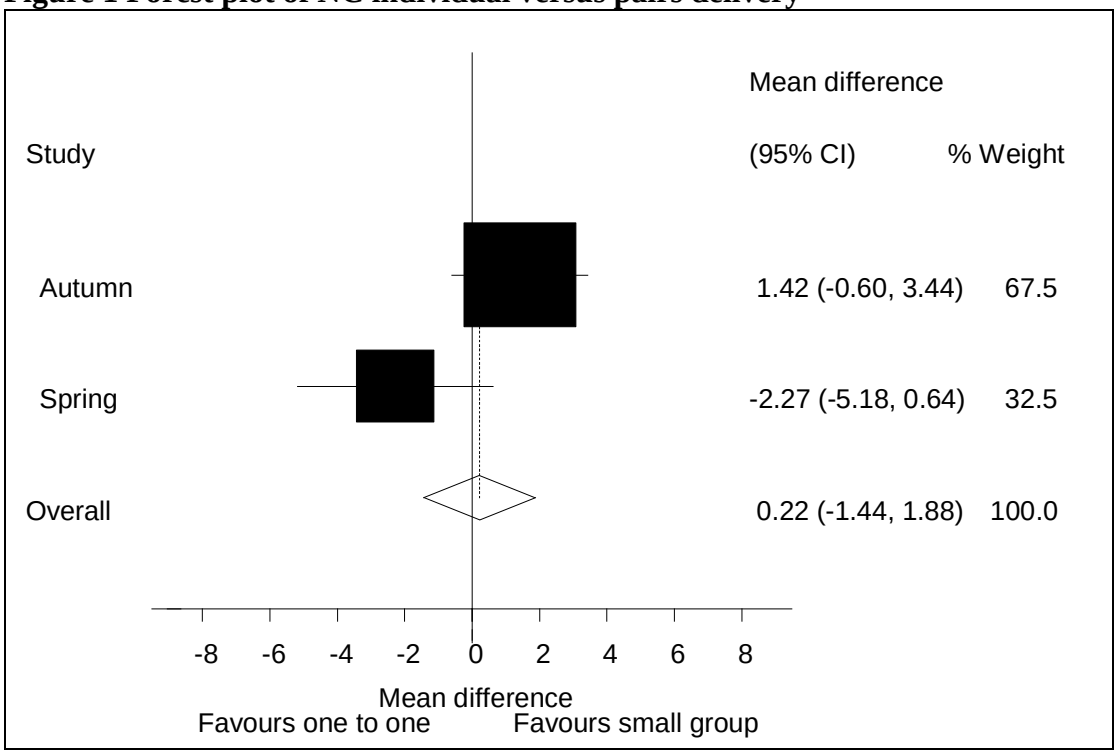


Figure 2 Forest plot of NC individual versus triplet delivery

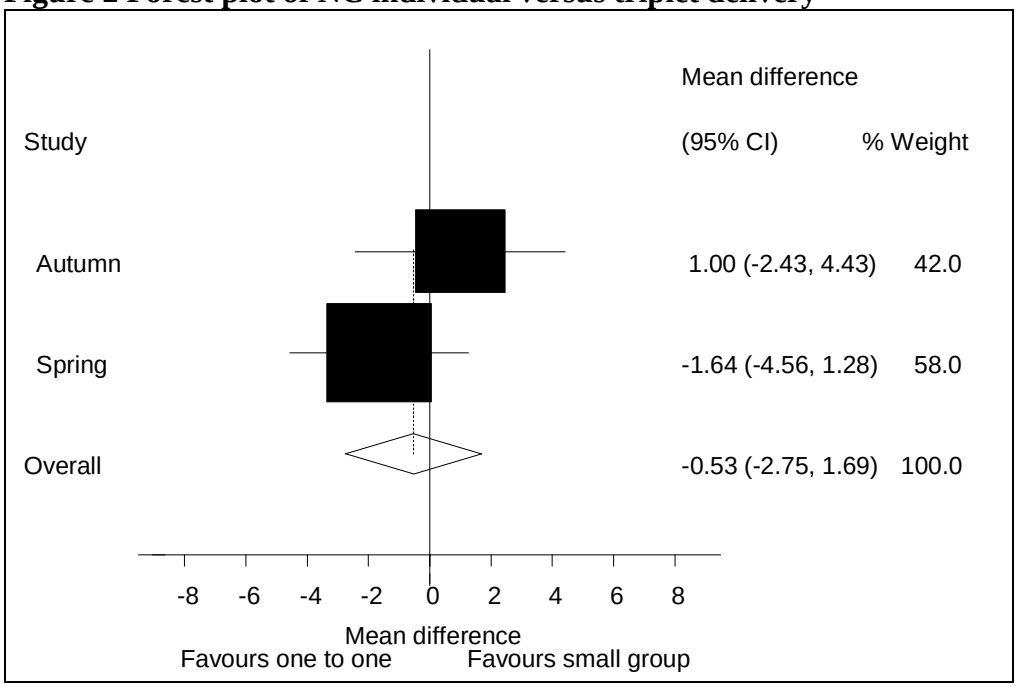


Table 3: Secondary outcomes

Pairs sub-trial

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)

Triplets sub-trial

Outcome	Individual	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.

Figure 3: Forest plot of individual vs. small group delivery of NC