Research in Mathematics Education
Vol. 15, No. 2, July 2013, xx-xx

Every Child Counts: testing policy effectiveness using a randomised controlled trial, designed, conducted and reported to CONSORT standards

Carole Torgerson[a]*, Andy Wiggins[b], David Torgerson[c], Hannah Ainsworth[c] and Catherine Hewitt[c]

[a]School of Education, Durham University, UK; [b]Centre for Evaluation and Monitoring, Durham University, UK; [c]Dept. of Health Sciences, University of York, UK

We report a randomised controlled trial evaluation of an intensive one-to-one numeracy programme – *Numbers Count* – which formed part of the previous government's numeracy policy intervention – *Every Child Counts.* We rigorously designed and conducted the trial to CONSORT guidelines. We used a pragmatic waiting list design to evaluate the intervention in real life settings in diverse geographical areas across England, to increase the ecological validity of the results. Children were randomly allocated *within* schools to either the intervention (*Numbers Count* in addition to normal classroom practice) or the control group (normal classroom practice alone). The primary outcome assessment was the Progress in Maths (PIM) 6 test from GL Assessment. Independent administration ensured that outcome ascertainment was undertaken blind to group allocation. The secondary outcome measure was the Sandwell test, which was not undertaken and marked blind to group allocation.
At post-test the effect size (standardised mean difference between intervention and control group) on the PIM6 was d=0.33 95% confidence intervals [0.12, 0.53], indicating strong evidence of a difference between the two groups. The effect size for the secondary outcome (Sandwell test) was d=1.11 95% CI [0.91, 1.31]. Our results demonstrate a statistically significant effect of *Numbers Count* on our primary, independently marked, mathematics test. Like many trials, our study had both strengths and limitations. We feel, however, due to our *a priori* decision to report these in an explicit manner, as advocated by the CONSORT guidelines, that we could maximise rigour (e.g., by using blinded independent testing) and report potential problems (e.g., attrition rates). We have demonstrated that it is feasible to conduct an educational trial using the rigorous methodological techniques required by the CONSORT statement.

Keywords: experimental design; mathematics education; CONSORT guidelines

*Corresponding author. Email: carole.torgerson@durham.ac.uk

ISSN 1479-4802 print/ISSN 1574-0178 online

DOI: 10.1080/xxxxx
http://www.tandfonline.com

Introduction

The randomised controlled trial (RCT) is the best available design for establishing effectiveness (Cook and Campbell 1979; Shadish, Cook, and Campbell 2002; Torgerson and Torgerson 2008). Nevertheless, the quality of RCTs varies, and a poor quality RCT may provide less reliable evidence than a well-designed and conducted non-randomised study. It is often difficult, however, to distinguish between rigorously or poorly designed and conducted RCTs unless the key methodological aspects that contribute to their robustness are well reported.

The problem of poorly reported RCTs is widely acknowledged in health care research (Schultz et al. 1995). A methodological comparison between reporting standards of RCTs in education and health care research found that educational trials were reported more poorly than health care trials (Torgerson et al. 2005). For instance, in a sample of more than 80 educational trials, none reported their rationale for sample size or whether randomisation was concealed and independent, and most did not report independent blinded outcome assessment but *did* report a number of outcomes without pre-specification of the main outcome. In response to acknowledgement of poor trial reporting in health care research, methodologists designed the Consolidated Standards for Reporting Trials (CONSORT) (Moher et al. 2010, Schultz et al. 2010). Most high profile medical journals and journals in other disciplines (e.g., all American Psychology Association journals and some education journals) now adhere to the CONSORT guidance. Although the CONSORT framework is easily adaptable to educational trials (see Torgerson and Torgerson 2008), it is not widely used in the reporting of educational field trials undertaken in the United Kingdom, although in the United States, the Institute of Education Sciences (http://ies.ed.gov/) recommends its use to researchers designing and reporting efficacy and effectiveness trials in educational research.

In this paper we present a randomised controlled trial evaluation of the effectiveness of an intensive one-to-one numeracy programme – *Numbers Count (NC)* – which is part of the previous governments' numeracy policy, *Every Child Counts (ECC)*. *ECC* was set up in 2007 as a partnership between Government, businesses and charities, and was administered by the Every Child a Chance Trust (ECaCT). The programme was funded by the Department for Children, Schools and Families (DCSF) and a number of charitable trusts, led by KPMG through ECaCT. From September 2011 the programme entered a three year transition phase. At the end of this period, the Department for Education (DE) will no longer administer the programme centrally, although schools will be free to buy into it themselves. The trial reported here is one aspect of the independent evaluation of the policy, which included an impact evaluation (three randomised controlled trials with an embedded economic evaluation) and a process evaluation of the implementation of the intervention. The evaluation was funded in 2009 by the Department for Children, Schools and Families (DCSF), and has previously been reported in detail in Torgerson et al. 2011a and Torgerson et al. 2011b. The technical report and accompanying appendices, including the trial protocol, can be located here: https://www.education.gov.uk/publications/RSG/SchoolsSO/Page8/DFE-RR091A.

The aim of this paper is to demonstrate that the key elements of the trial, including key methodological aspects as well as the principal results, can be reported in a concise fashion within a journal, without the need for a lengthy report. Therefore, all the crucial elements a reader might wish to know about the design and conduct of the trial are contained within this paper, which allows an objective judgement to be made about its methodological quality. The 22-item CONSORT checklist has been included as Appendix A, and a CONSORT flow diagram has been included as Figure 1 in order to demonstrate that it is

possible to design, conduct and report an educational trial using the same quality standards that are routinely expected in health care trials.

Background

In 2011 in the UK the number of 11 year-olds gaining level 4 and above (expected levels) at key stage 2 mathematics was 80%, having risen from 59% in 1998. A recent independent review for the UK government of mathematics teaching in primary schools, including *ECC*, noted that about 5% of lower attaining pupils at age 11 go on to leave secondary education with no qualification at all in mathematics (Williams 2008).

The development of the *ECC* programme was supported by the then-Labour government to address underachievement in numeracy skills in primary schools. *Every Child Counts* includes three 'waves' of mathematics instruction and intervention: wave 1 (quality classroom teaching for all), wave 2 (small group additional intervention for children just below national expectations) and wave 3 (individual intervention with a trained specialist teacher). *Numbers Count* provides the wave 3 intensive one-to-one intervention for those children identified as lowest attaining 6–7 year old children at risk of failing to thrive in numeracy (Edge Hill University 2008). It is a 12-week programme consisting of daily 30 minute sessions delivered by specially trained teachers, and takes place during normal lesson time. It was specifically designed to help children to develop their knowledge and understanding of number, through a comprehensive diagnostic assessment of each child's strengths and weaknesses, core learning objectives, and guidance for teachers on lesson structure and key teaching approaches. *Numbers Count* was an expensive centrally funded intervention; therefore it was essential that it was robustly and independently evaluated.

In order to obtain reliable evidence of the effectiveness of *NC* compared with normal classroom practice, the UK government funded us to undertake an independent evaluation using a randomised controlled trial design.

An RCT is superior to an evaluation using a single group pre- and post-test design. Results from evaluations using this (pre-experimental) design are likely to 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 whereby 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 was also crucial that such a control group was prospectively assembled through the process of random assignment, otherwise bias could have been introduced (Torgerson and Torgerson 2008). Such bias can either underestimate or overestimate the effectiveness of an intervention. For instance, if pupils who received *NC* in our evaluation were selected on the basis of a low score on a test, then regression to the mean effects would have ensured an exaggerated improvement compared with children in the control group who scored higher on the pre-test and did not receive the intervention. Alternatively, teachers who select children for an intervention may, consciously or unconsciously, select children that they think will do especially well; and comparing their performance with other children, even with similar test scores, will produce a biased result. Random allocation ensures that biases due to temporal changes, regression to the mean, or pupil selection are absent from effect size estimates.

Although random allocation eliminated selection bias, the potential for a number of post-randomisation biases to be introduced after random allocation was minimised through design. For example, it was important that tests undertaken post-randomisation were administered and marked by personnel who were *blinded* or *masked* to the membership of the intervention and control groups. This was to avoid conscious or unconscious effects by testers who may have had a desire to ensure that the intervention children performed to the best of their ability.

Design and methods

We used a pragmatic randomised design (Torgerson and Torgerson 2007). This design was used to evaluate the intervention in real life classroom and school settings in diverse geographical areas across England to increase the ecological validity and the extent to which the results could be generalised to other schools not included in the trial. In this design, children and their parents identified by the schools as eligible to receive *NC* were recruited and consented in the first two weeks of term in September 2009. The children were randomly allocated *within* schools to either the intervention or control group. The intervention group received *NC* in the autumn term (in addition to normal classroom maths teaching), whilst the control group received normal classroom maths teaching alone during the autumn term, and were placed on a waiting list to receive *NC* in addition in the spring or summer terms. The decision to use a delayed treatment (waiting list) design with unequal allocation was a pragmatic one, and taken to enable all 12 children eligible to receive *NC* in each school to be included in the randomised comparison at post-test. This design, therefore, gave the optimum number of children included in the trial.

The primary outcome assessment was the Progress in Maths (PIM) 6 test from GL Assessment (Clausen-May et al. 2004). The assessment covers a wide range of mathematical

skills (number, shape and data handling) and is broader than the scope of the skills covered by *NC*. It should be noted, however, that the primary justification for the *ECC* programme was to raise mathematical achievement generally, and *NC* sought to do this by concentrating on number, on the assumption that other areas of mathematics would also improve (Edge Hill University 2008). It follows, therefore, that a more general test of mathematical knowledge, as opposed to one that just focuses on number, is an appropriate primary outcome. PIM 6 is not a programme-inherent measure, which means that the children were not directly taught the concepts included in the outcome measure. Independent administration meant that the evaluators could ensure that outcome ascertainment was undertaken blind to group allocation.

The baseline test for the PIM 6 was the Sandwell test (Arnold et al. 2011); we used this to increase the power and precision of the study. We also used the Sandwell test as a secondary outcome measure at post-test. This 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 *NC*. The assessment covers national curriculum skills from P6 to level 2a, focuses on number and largely coincides with the underlying approach of *NC*. However, in our trial, the Sandwell test was not undertaken independently of the implementation of the *NC* programme. Therefore, we, as the evaluators, could not ensure that the people administering and marking the test at post-test did not know whether the children were in the *NC* intervention group or in the control group. The post-test Sandwell test results should be treated with appropriate caution.

Randomisation

Once baseline testing was completed, the children were randomly allocated within schools using an independent, concealed randomisation process which was undertaken by the York Trials Unit. An independent data manager, from the York Trials Unit, wrote a software

programme that randomised participants in exact numbers into three groups: *NC* delivered in the autumn term (i.e., the intervention group) and *NC* delivered in either the spring or summer term (i.e., the control group). Thus, from a block of 12 children from each school, 4 were randomly allocated into each group: no other stratification variable was used. The investigators had no role in the randomisation procedure. The use of a secure, third party allocation system though the York Trials Unit ensured that the random allocation was concealed and independent of the developers and implementers of the intervention, and that it could not be tampered with in ways which have been shown to be problematic in some health care trials (Schulz 1995).

Blinded assessment

To reduce the problem of potential ascertainment bias, we used independent testers for the PIM 6 test who did not know whether the children they tested were in the intervention or the control group. The tests were also marked blindly.

All children were tested in January 2010, after the intervention group had received *NC* and before the control group received *NC*, using the PIM 6 mathematics test, the primary outcome measure. All children were also tested at the same time using the secondary outcome mathematics test, the Sandwell test.

Sample size calculation for the PIM 6 test

The power calculations were based upon the following. We wanted to detect a difference in PIM 6 of 0.25 standard deviations between the intervention and control groups. We also assumed a pre-test post-test correlation of at least 0.70 (i.e., the Sandwell test would correlate by at least 0.7 with the PIM 6). To have at least a 95% chance of observing such a difference we needed approximately 600 children in our sample given a randomisation ratio of 2:1 (at

the end of the first term generally 8 children were in the control group and 4 were in the intervention group). Given that *NC* was offered to 12 children in most of the sample schools, we needed to recruit 50 schools.

Statistical analysis

We prepared a statistical analysis plan before the data were analysed. The primary analysis compared mathematics attainment on the PIM 6 test of the intervention children receiving *NC* in the autumn term with the control children who had not yet received *NC* and were allocated to receive *NC* in the spring or summer terms. The secondary analysis compared mathematics attainment on the Sandwell test of the intervention children receiving *NC* in the autumn term with the control children who had not yet received *NC* and were allocated to receive *NC* in the spring or summer terms. The analysis of the PIM 6 test was conducted on an intention-to-treat basis, which provides the most useful indication about the impact of the programme. Intention-to-treat analysis means that any children who crossed over from either study arm (i.e. dropped out or received *NC* at a different time) were analysed as per their randomised 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 descriptively. The scores on the PIM 6 were summarised descriptively (means and standard deviations) by allocated group (intervention and control). 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). This was because, even though the children were allocated individually, they were grouped within schools to receive the intervention. The outcome modelled was the PIM 6 score, and the model included age, gender, free school meal status, Sandwell test score (pre-test) and group allocation. This analysis was repeated for the secondary outcome, which was the Sandwell test.

Results

Data from 44 schools and 522 children were included in the analysis for this report. The progress and attrition of schools and children through the trial is shown in the following CONSORT diagram, Figure 1.

[FIGURE 1 HERE]

There were relatively few protocol deviations; and we adhered to an intention-to-treat data analytic plan.

Table 1 summarises characteristics of all children included in the trial by the term of delivery. 18 children were randomised to receive *NC* in the spring or summer terms only for pragmatic reasons, and have been excluded from the summaries below. As expected, randomisation resulted in all groups having similar characteristics.

[TABLE 1 HERE]

Between randomisation to intervention group or control group and assessment on the PIM 6 test, approximately 86 (17%) of children were lost to follow-up or were withdrawn. There were a number of reasons for those lost to follow-up, including absence from school during the testing and the bad, snowy weather in the UK in January 2010 when the post-tests were undertaken. However, we do not believe that the absence of these children is likely to have introduced bias, as the proportion missing from each group was similar: 31 (18%), 24 (14%) and 31 (19%) for groups allocated to autumn, spring and summer respectively, and there did

not appear to be any systematic reasons for the drop-out that would have been related to the group to which the children had been allocated.

In Table 2 we show the main results. The mean PIM 6 mathematics test score for the intervention children was 15.8 (SD 4.9) and the mean score for the control children was 14.0 (SD 4.5). The effect size (Cohen's d) was d=0.33 95% CI [0.12, 0.53], with confidence intervals a long way from zero, indicating strong evidence of a difference between the two groups, adjusted B 1.47 95% CI [0.71, 2.23] p<0.001.

[TABLE 2 HERE]

We estimate that, on average, *NC* produced an additional improvement of 7 weeks in numeracy skills, as measured by the PIM 6 mathematics test, compared with usual teaching. In other words, in a 12-week term the children in the intervention group improved by 19 weeks compared with the children in the control group, who improved by 12 weeks.

The secondary outcome measure was the Sandwell test, which was undertaken and marked by *NC* teachers, class teachers or teaching assistants, who were not blind to group allocation. The effect size for this measure was much larger at d=1.11 95% CI [0.91, 1.31].

Table 3 shows that the children improved their Sandwell test scores once they received *NC*. By July, when all children had received *NC*, they were all performing at a similar level in numeracy, as measured by the Sandwell test.

[TABLE 3 HERE]

Discussion

We report the results of a randomised controlled trial comparing one-to-one *NC* teaching with normal classroom practice using CONSORT criteria. Because we adhered to this guidance, we argue that readers can assess the rigour of our design and understand its limitations.

In summary, our results demonstrate a statistically significant effect of *Numbers Count* on our primary, independently marked, mathematics test. The effect size of d=0.33 is reasonable for a pragmatic field trial. We estimate that this translates into an average of 7 additional weeks' progress for children in the intervention (*NC*) group over the course of a 12-week term, or 19 weeks' progress in numeracy for children receiving *NC* compared with 12 weeks' progress for children not receiving *NC*.

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 average score of the control group, the results are consistent with an extra 12–16% of the children in the intervention group getting a score higher than the average score of the control group.

The trial design has a number of key strengths. The randomisation 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. We also present our results as differences in means with 95% confidence intervals. The use of confidence intervals allows us to understand the likely range of intervention effect. CONSORT recommends the use of confidence intervals rather than p values when presenting the results from pragmatic field trials. Historically, few RCTs in education have presented confidence intervals; typically there is an emphasis on p values when reporting trials

(Torgerson et al 2005). Finally, we, the evaluators, were independent of all parties with a potential interest in the outcomes of the trial.

The trial does, however, have a number of limitations. As described in the CONSORT flow diagram, we had a significant number of children not taking the PIM 6 Mathematics test. However, because the number of children missing was virtually identical between the intervention and control groups, and because there did not appear to be any systematic reason for the drop-out, we do not think any significant bias was introduced. In Table 4 we show the baseline characteristics for all children in the trial, showing separately those pupils included in the primary analysis and those with missing primary outcome data.

[TABLE 4 HERE]

Our actual sample size was somewhat lower than we had anticipated (418 rather than 600); however, the effect size we observed (d=0.33) was somewhat greater than anticipated in the original sample size calculation. Consequently, there was little loss of power in our study. Although our secondary outcome measure of mathematical achievement showed a much larger difference (i.e., >1 standard deviation), we must be cautious when interpreting this difference, as it was not marked blindly and is a treatment inherent measure, both of which factors will tend to exaggerate the effectiveness of the intervention. Therefore, the higher effect size for this measure compared with the effect size for the PIM 6 measure may be due to a conscious or unconscious tendency, by the markers, to award higher marks to intervention group children when undertaking the Sandwell tests, or because it was a treatment inherent measure or a combination of both.

The design did suffer from other limitations, which were partly due to the funders requiring us to use a short-term waiting list design which prevented us from looking at the

longer-term effects of *NC*. Ideally, a cluster randomised design utilising a longer term follow-up would be necessary to see whether or not the intervention could have 'washed out' over time, with the children in the control group catching up using normal classroom teaching. Furthermore, we could not disentangle the effect of one-to-one teaching *per se* from *NC*, which is delivered individually. Consequently, it may be that offering a different one-to-one mathematics intervention could have had similar effects.

Conclusions

In summary, our data demonstrate that the short-term impact of *Numbers Count* is positive. Our results demonstrate a statistically significant effect size of d=0.33 of *NC* on the primary mathematics test. The effect size translates into an average of 7 additional weeks' progress for children in the intervention group over the course of a 12-week term.

Like many trials, our study had both strengths and limitations. We feel, however, due to our *a priori* decision to report these in an explicit manner, as advocated by CONSORT, that we limited weaknesses where we were able (e.g., using blinded independent testing) and reported potential problems where we could not (e.g., attrition rates). This allows readers to evaluate the scientific rigour of our study, and make better informed decisions in terms of policy and future research. Finally, we have demonstrated that it is feasible to conduct an educational trial using the rigorous methodological techniques required by the CONSORT statement.

Acknowledgements

We acknowledge all members of the ECC evaluation team: Carole Torgerson, Andy Wiggins, David Torgerson, Hannah Ainsworth, Patrick Barmby, Catherine Hewitt, Karen Jones, Vivien Hendry, Mike Askew, Martin Bland, Rob Coe, Steve Higgins, Jeremy Hodgen, Charles Hulme and Peter Tymms. The trial described in this article was funded by the Department for Children, Schools and Families (DCFS). We also acknowledge Durham University and the University of York for additional funding to support the trial. The corresponding author was employed by the University of York when the field work for the trial was undertaken.

References

Arnold, C., P. Bowen, M. Tallents, B. Walden, and Sandwell Inclusion Support Service. 2011. *Sandwell early numeracy test – revised (SENT-R).* GL Assessment.

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

Cook, T.D., and D. Campbell. 1979. *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.

Lipsey, M.W., and D.B. Wilson. 1993. The efficacy of psychological, educational and behavioral treatment: Confirmation from meta-analysis. *American Psychologist* 48: 1181–209.

Moher, D., S. Hopewell, K.F. Schulz, V. Montori, P.C. Gøtzsche, P.J. Devereaux, D. Elbourne, M. Egger, D.G. Altman, for the CONSORT Group. 2010. Explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. BMJ 2010; 340: c869.

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

Schultz, K.F., I. Chalmers, R.J. Hayes, and D.G. Altman. 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: 408–12.

Schulz, K.F., D.G. Altman, D. Moher for the CONSORT Group. 2010. Updated guidelines for reporting parallel group randomised trials. *Annals of Internal Medicine* 2010: 152.

Shadish, W.R., T.D. Cook, and D. Campbell. 2002. Experimental and quasi-experimental designs for generalised causal inference. Boston: Houghton Mifflin.

Torgerson, C.J., and D.J. Torgerson. 2007. The need for pragmatic experimentation in educational research, *Economics of Innovation and New Technology* 16: 323–30.

Torgerson, D., and C. Torgerson. 2008. *Designing and running randomised trials in health, education and the social sciences*. Hampshire and New York: Palgrave Macmillan.

Torgerson C.J., D.J. Torgerson, Y.F. Birks, and J. Porthouse. 2005. A comparison of randomised controlled trials in health and education. *British Educational Research Journal* 31: 761– 85.

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

2011a. *Every Child Counts: The independent evaluation executive summary*. Department for Education.

Torgerson, C., A. Wiggins, D.J. Torgerson, H. Ainsworth, P. Barmby, C. Hewitt, K. Jones, V. Hendry, M. Askew, M. Bland, R. Coe, S. Higgins, J. Hodgen, C.Hulme and P. Tymms. 2011b. *Every Child Counts: The independent evaluation – Appendices.* Department for Education.

Williams, P. 2008. *Independent review of mathematics teaching in early years settings and primary schools*. Department for Children, Schools and Families (DfCSF).

Appendix A. CONSORT checklist (Schultz et al. 2010; Moher et al. 2010)

Section/Topic	Item No	Checklist item	Reported on page No
Title and abstract			
	1a	Identification as a randomised trial in the title	1
	1b	Summary of trial design, methods, results, and conclusions	1
Introduction			
Background and objectives	2a	Scientific background and explanation of rationale	5–6
	2b	Specific objectives or hypotheses	3
Methods			
Trial design	3a	Description of trial design (such as parallel, factorial) including allocation ratio	6–7
	3b	Important changes to methods after trial commencement (such as eligibility criteria), with reasons	N/A
Participants	4a	Eligibility criteria for participants	4
	4b	Settings and locations where the data were collected	6
Interventions	5	The interventions for each group with sufficient details to allow replication, including how and when they were actually administered	4
Outcomes	6a	Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed	7–8
	6b	Any changes to trial outcomes after the trial commenced, with reasons	N/A
Sample size	7a	How sample size was determined	9
	7b	When applicable, explanation of any interim analyses and stopping guidelines	N/A
Randomisation:			
Sequence generation	8a	Method used to generate the random allocation sequence	8
	8b	Type of randomisation; details of any restriction (such as blocking and block size)	8
Allocation concealment mechanism	9	Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned	8
Implementation	10	Who generated the random allocation sequence, who enrolled participants, and who assigned participants to interventions	8
Blinding	11a	If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how	7–8
	11b	If relevant, description of the similarity of interventions	N/A
Statistical methods	12a	Statistical methods used to compare groups for primary and secondary outcomes	9–10
	12b	Methods for additional analyses, such as subgroup analyses and adjusted analyses	9–10
Results			
Participant flow (a diagram is strongly recommended)	13a	For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome	10–11

	13b	For each group, losses and exclusions after randomisation, together with reasons	10–11
Recruitment	14a	Dates defining the periods of recruitment and follow– up	6, 8–9
	14b	Why the trial ended or was stopped	N/A
Baseline data	15	A table showing baseline demographic and clinical characteristics for each group	11
Numbers analysed	16	For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups	10–11, 9
Outcomes and estimation	17a	For each primary and secondary outcome, results for each group, and the estimated effect size and its precision (such as 95% confidence interval)	11, 12
	17b	For binary outcomes, presentation of both absolute and relative effect sizes is recommended	N/A
Ancillary analyses	18	Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory	N/A (in full report)
Harms	19	All important harms or unintended effects in each group (for specific guidance see CONSORT for harms)	N/A
Discussion			
Limitations	20	Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses	14
Generalisability	21	Generalisability (external validity, applicability) of the trial findings	6
Interpretation	22	Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence	12–15
Other information			
Registration	23	Registration number and name of trial registry	N/A
Protocol	24	Where the full trial protocol can be accessed, if available	3
Funding	25	Sources of funding and other support (such as supply of drugs), role of funders	16

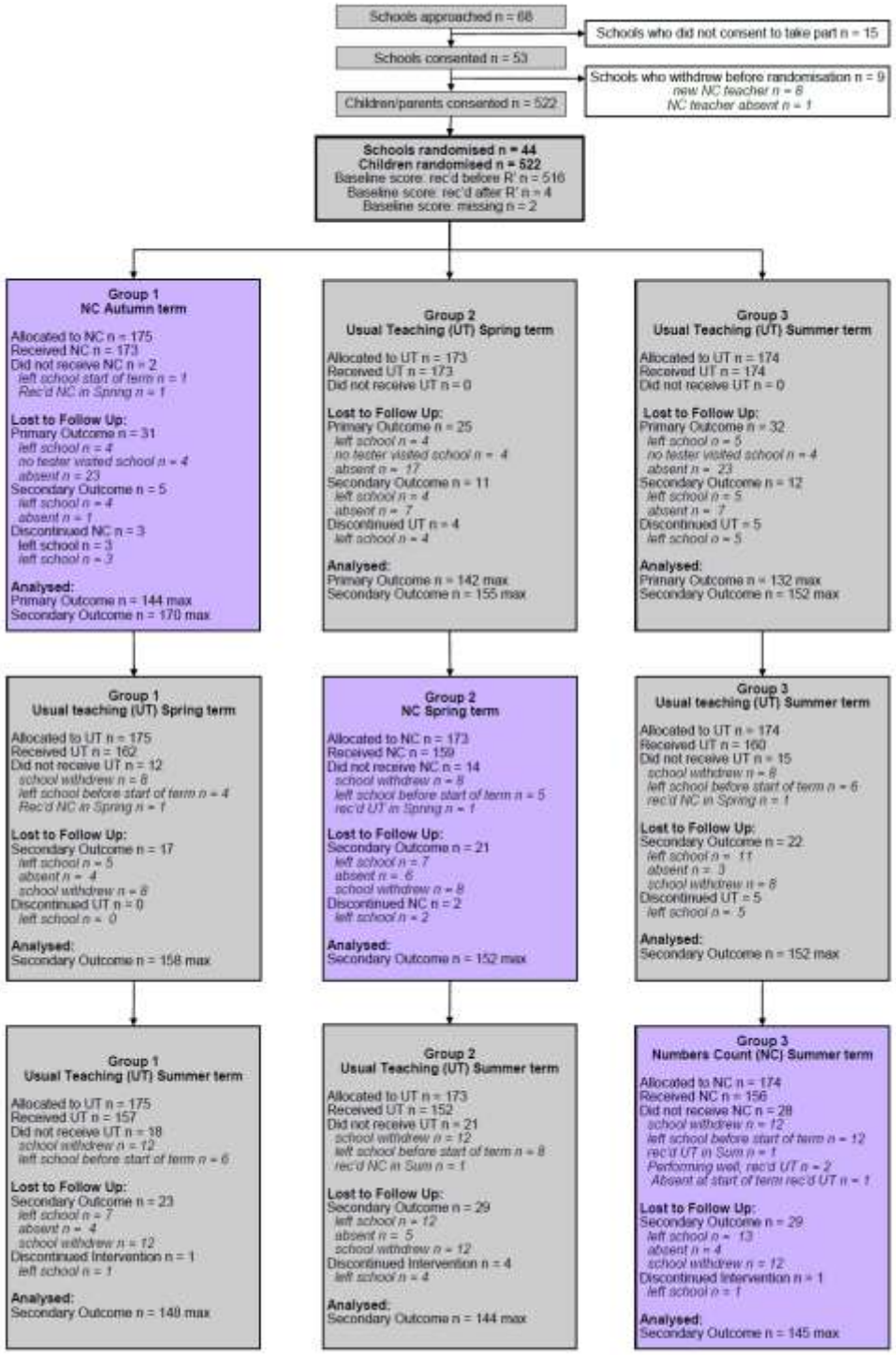

Figure 1. CONSORT diagram.

Table 1. Baseline characteristics.

Characteristics of all children included in Trial 1	**Intervention group (n=175)**	**Control group (n=329)**
Age: mean (standard deviation)	6.4 (0.3) [n=173]	6.4 (0.3) [n=327]
Sandwell A Mathematics test score in Sept. 2009: mean (standard deviation)	28.2 (8.4) [n=174]	26.8 (8.5) [n=328]
Children who received free school meals: number (%)	86 (50.9) [n=169]	139 (43.9) [n=317]
Gender (females): number (%)	69 (39.7) [n=174]	137 (41.8) [n=328]

Note. Excludes children unable to be randomised to the autumn term

Table 2. Summary of primary outcome measure.

	PIM 6	
	B	95% CI
Randomised group	1.47	[0.71, 2.23]*
Baseline Sandwell A	0.41	[0.37, 0.46]*
Free school meals	-0.33	[-1.00, 0.33]
Gender	-0.45	[-1.09, 0.19]
Age	-0.70	[-1.97, 0.58]
Constant	7.69	[-0.26, 15.64]
R^2	0.56	
F	66.32*	
Effect size (Cohen's d) with 95% confidence intervals (CI)	d=0.33 (0.12 to 0.53)	

Note. N = 409. CI = confidence interval. Analyses were adjusted for the clustering within schools. Analyses excluded children who could not be randomised to autumn term.
*p<0.001

Table 3. Sandwell mathematics test scores of all children**.**

Mathematics assessment	**Children who received Numbers Count in the autumn term**		**Children who received Numbers Count in the spring term**		**Children who received Numbers Count in the summer term**	
	N	**Mean score (SD)**	**N**	**Mean score (SD)**	**N**	**Mean score (SD)**
Sandwell A (Sept 2009)	174	28.2 (8.4)	165	26.7 (8.3)	163	27.0 (8.7)
Sandwell B (Jan 2010)	170	45.0 (11.1)	155	32.3 (9.9)	152	32.7 (10.6)
Sandwell A (Apr 2010)	158	48.7 (10.6)	147	48.2 (12.1)	144	37.0 (11.0)
Sandwell B (Jul 2010)	152	52.8 (11.4)	139	51.9 (13.1)	137	50.9 (12.5)

Note. Excludes children unable to be randomised to the autumn term.

Table 4. Baseline characteristics: primary outcome data and missing outcome data.

Characteristics of all children included in trial	**Pupils with primary outcome data (n=418)**	**Pupils with missing primary outcome data (n=86)**
Age: mean (standard deviation)	6.4 (0.3) [n=418]	6.4 (0.3) [n=82]
Sandwell A mathematics test score in Sept. 2009: mean (standard deviation)	27.5 (8.4) [n=417]	26.4 (8.9) [n=85]
Children who received free school meals: number (%)	189 (46.1) [n=410]	36 (47.4) [n=76]
Gender (females): number (%)	178 (42.6) [n=418]	28 (33.3) [n=84]

Note. Excludes children unable to be randomised to the autumn term.