CfBT Education Trust

Evidence that counts – what happens when teachers apply scientific methods to their practice

Twelve teacher-led randomised controlled trials and other styles of experimental research

Commentary by Richard Churches and Tony McAleavy

Providing education for public benefit worldwide

Welcome to CfBT Education Trust

CfBT Education Trust is a top 30* UK charity providing education services for public benefit in the UK and internationally. Established over 40 years ago, CfBT Education Trust has an annual turnover exceeding £100 million and employs more than 2,000 staff worldwide. We aspire to be the world's leading provider of education services, with a particular interest in school effectiveness.

Our work involves school improvement through inspection, school workforce development and curriculum design for the UK's Department for Education, the Office for Standards in Education, Children's Services and Skills (Ofsted), local authorities and an increasing number of independent and state schools, free schools and academies. We provide services direct to learners in our schools and in young offender institutions.

Internationally we have successfully implemented education programmes for governments in the Middle East, Sub-Saharan Africa and South East Asia and work on projects funded by donors such as the Department for International Development, the European Commission, the Australian Department of Foreign Affairs and Trade, the World Bank and the US Agency for International Development, in low- and middle-income countries.

Surpluses generated by our operations are re-invested in our educational research programme.

Visit www.cfbt.com for more information.

**CfBT is ranked 27 out of 3,000 charities in the UK based on income in Top 3,000 Charities 2010/11published by Caritas Data*

Contents

About the authors of the commentary

Richard Churches

Richard Churches is Principal Adviser for Research and Evidence Based Practice at CfBT Education Trust. He has worked as a government adviser on multiple policy initiatives in the UK and abroad, recently leading the Closing the Gap: Test and Learn programme which implemented randomised controlled trials in collaboration with 200 teaching schools and 800 trial site schools. His doctoral research was experimental and looked at the hypnotic effects of charismatic oratory on the structure of consciousness. Whilst conducting his research at the University of Surrey he was awarded a Pearson Education prize, won the Vitae Three Minute Thesis Competition and following the successful defence of his thesis was nominated as Post-Graduate Student of the Year. Richard is co-recorder for the British Science Festival Education Section with Eleanor Dommett (Institute of Psychiatry, Psychology and Neuroscience, King's College London) with whom he recently wrote the book *Teacher-led research: how to design and implement randomised controlled trials and other forms of experimental research* (Crown House Publishing).

Tony McAleavy

Tony McAleavy is CfBT Education Trust's Research and Development Director, with corporate oversight of the educational impact of all CfBT's activities and CfBT's public domain research programme. CfBT's practice-based research helps to inform and shape education policy in the UK and worldwide. The aim is to ensure that our investment has a direct and powerful impact on beneficiaries via practitioners and policymakers. Tony has worked extensively on school reform in many countries, particularly in the Middle East. He has an MA in Modern History from St John's College, University of Oxford.

The teacher-researchers

Kate Andrews
Simon Andrews
David Ashton
Timm Barnard-Dadds
Sarah Baugh-Williams
Bridie Bear
Ceri Bibby
Wendy Blyth
Allison Davies
Rachel Elphick
Dean Flood
Emmet Glackin
Gavin Jones
Graeme Jones
Daniel Lear
Matthew Maughan
Jess Moore
Charlotte Morris
Theresa Peacock
James Siddle
Alison Turner
Rob Wilson

Acknowledgements

This piece of work would not have been possible without contributions from a wide range of people. CfBT Education Trust would like to thank the 22 teachers (listed) who kindly gave permission to include their poster reports in this publication and Marian Gould and Richard Warenisca for their help in producing this publication. We would also like to acknowledge the considerable support given by Paul Booth, Anna Brychan and Delyth Balman of the National Support Programme in Wales;[1] and their facilitators who assisted the teachers during the training programme and with the implementation of their research (Karen Norton, Kate Andrews, Finola Wilson and Jane Miller).

The research projects undertaken by Kyra Teaching School Alliance were made possible through a research grant from the Department for Education/National College for Teaching and Leadership, Closing the Gap: Test and Learn programme.[2] On this programme, the first experimental research design training for teachers took place with the notable encouragement of Robin Hall and Juliet Brookes, resulting in over 50 teacher-led Randomised Controlled Trials (RCTs) being initiated. In addition, we would like to acknowledge the role of Rebecca Annand (Infinite Learning) who supported the first international delivery of training for teachers in this area as part of the 2014-2015 BSME (British Schools in the Middle East) programme of CPD events in Dubai. Infinite Learning is offering an extended version of the programme in Dubai for the academic year 2015/2016 (see: www.infinitelearning.ae). Finally, the teachers would not have been able to analyse their results without the combined work of Richard Churches, Jan Lawrance and Mick Blaylock who together authored the Excel worksheet programmes that allowed for the use of the inferential tests used in the teachers' reporting.

[1] The National Support Programme was funded by the Welsh Government and implemented on their behalf by CfBT Education Trust between 2013 and 2015; providing support for all 1,608 schools in Wales.
[2] Closing the Gap: Test and Learn was delivered by CfBT Education Trust on behalf of the National College for Teaching and Leadership, between 2013 and 2015, in partnership with the University of Oxford, CUREE (Centre for the Use of Research and Evidence in Education) and Durham University.

Executive summary

This publication contains 12 (A3 open-out) poster-style reports of teacher experimental research. The style of presentation parallels the type of preliminary reporting common at academic conferences and postgraduate events. At the same time, it aims to act as a form of short primer to introduce teachers to the basic options that there are when conducting this type of research. This said, the determined teacher-researcher would be advised to extend their reading further to full-scale textbooks such as Richard Churches and Eleanor Dommett's *Teacher-led research: how to design and implement randomised controlled trials and other forms of experimental research* which clearly explains how to choose and conduct the right statistical tests. CfBT Education Trust also offers an in-depth training programme over two days with access to statistical software and support for teachers when conducting and writing up their research. Most of the teachers in this report followed this programme in Wales, or in one case Dubai. Others, associated with Kyra Teaching School Alliance (led by Mount Street Academy, part of the CfBT Schools Trust), attended the Department for Education/National College for Teaching and Leadership training that Richard Churches delivered as part of the Closing the Gap: Test and Learn programme in England.

The reports in this publication demonstrate the potential of practising teachers to carry out research which applies a scientific method, both in terms of the generation of school-based local and contextual evidence and with regard to the development of a research-engaged, research-informed and research-led teaching profession. They also illustrate how poster-style reporting can offer an engaging and immediate way of understanding a piece of education research of this type, how the research was conducted and its implications for other teachers. The report was previewed at the 2015 British Science Festival within a session delivered by Richard Churches, Tony McAleavy and Michael Latham.

To find out more about the CfBT Education Trust training programme which supports teachers and schools to conduct experimental research contact: rchurches@cfbt.com.

1. Introduction

1.1 | Teaching as a research-informed, research-engaged and research-led profession; the importance of the emerging approaches documented in this report

Moving teaching into a space in which evidence is front and centre of practice is without doubt a good thing. Indeed, such a move has the potential to make a difference in a number of ways, not least of which include:

- the development of more powerful professional development
- improved teaching techniques
- better whole-school decision making
- improved outcomes for students.

However, this shift is not an easy one. Indeed, developments over the past five years and the emerging mindset that has led to the sort of approaches documented in this report throw into sharp relief a number of issues.

Education research receives a tiny fraction of the funding available for medical research. Alongside this, most publicly-funded education research initially appears in journals that teachers cannot access. It is also the case that very little repetition in education research takes place to check whether the findings replicate. This picture is compounded by the fact that pedagogy in non-core subjects has been subject to very little rigorous research, with much education research being context-specific and not easily able to be generalised or transferred.

Two problems appear at this point to hold us back. Primarily, researchers themselves disagree about the applicability of the evidence, with much of the required evidence (and clarity over what is meant by an intervention) nowhere to be found. On another level, although there is a substantial body of research in some key areas (for example, the characteristics of schools that achieve good results, promising assessment techniques, aspects of special needs etc.), this is not enough. Rather (and accepting that it would be unprofessional to ignore this evidence) it still needs to be contextualised for each school. There is, however, light at the end of the tunnel and as this report demonstrates, schools can carry out their own rigorous local investigations. Not only that, teachers can carry out such studies in a way that can be replicated by others, and which can be scaled up to achieve larger sample sizes through collaboration.

The 12 examples presented also show the value of going beyond being just research-engaged and research-informed, to a third, much more important level of practice, as ‘research-led’ organisations in which evidence then informs the strategic decisions that are made within the school and the next steps in the school improvement process. For such a model to work it will need to be underpinned by a number of features:

- using enquiry to do the core business of school improvement
- embedding research as the collective work of a professional learning community

- strategic leadership and careful tailoring to school circumstance
- using enquiry as the engine of a self-improving school system
- making possible disciplined innovation through rigorous local enquiry.

As well as presenting evidence to show that the above is indeed possible, the studies that follow exemplify how experimental research methods can help to provide a focal point for such developments. First, and foremost perhaps, is the value of training teachers in a rigorous research method (including the correct use of research statistics) and then letting the teachers test what they think to be important. After so much evidence has pointed to effective school improvement being a balance between high accountability and high autonomy, it seems nothing more than 'a given' that where research evidence is concerned the same principles should apply. Namely, that teachers should not just be leading the creation of evidence; but that they should be doing so using structures and methods that have the highest levels of accountability built into them, paralleling the approaches long adopted by medicine and clinical practice.

Secondly, it is clear that there is still much to learn with regard to the application of experimental research methods in an education context. For where many of the approaches so far adopted have attempted to conduct only one style of study (typically one with just a control and intervention and comparing the difference between groups over six months) there are many other forms and approaches (common in psychology) that have application and value. At the same time, where a number of large-scale six-month-long trials have failed to yield significant findings, this is far from the case with the types of designs illustrated here. In contrast, the studies (and the other 70 that we have been involved with so far), which are more akin to the sort of tighter controlled approaches that again are to be found in laboratory-style psychology research, appear to have no problem identifying significant results in at least 50–60% of cases.

Thirdly, and by extension, science is not and has never been about conducting a single large-scale piece of research. The reason for this is very simple. If 100 studies are conducted then by definition it is likely that 95 of those studies will show one thing whilst the other five will show the opposite. This is because chance always plays a part in quantitative research – the reason why science always deals in probabilities and never truths. From enough replicated trials can eventually emerge a theory and, with enough time and replication, a theory that seems unlikely to be overturned. Adopting such an approach therefore demands that, alongside teaching and using scientific method, we look to where the capacity for such endeavour will eventually come. The idea that this is possible from just a few large-scale expensive trials (as important as these may be alongside teacher-led research) seems incomprehensible. Rather the answer surely is to build the capacity for the teaching profession itself to lead the way. For, paralleling the way that science grew and developed in the 18th and 19th centuries (through the endeavours of individuals), if only 10% of teachers in England conducted one robust randomised controlled trial over a five-year period we would have 40,000 trial results. The body of knowledge that these '40,000 (educational) Faradays' could provide, from their 'home' laboratories, would rival anything in the natural sciences. Of course, such a vision raises as many questions as it does answers; not least of which is whether we can ever reach this level of engagement. However, the enthusiasm for this type of research – we have so far observed – leaves no doubt in our minds that this area will grow and develop over the next decade.

1.2 | Understanding some important concepts before you read the reports

Before you read this report there are two concepts you need to familiarise yourself with. These are the notions of effect size and significance. Effect size is a measure that tells you the strength and direction of any change that has taken place. The two most common effect sizes you will come across are d and r (but there are others). For example, d = 0.5 is a moderate positive effect (i.e. there has been a positive shift in the spread and central tendency (or average) of your results that across most scientific studies would represent a moderate level of change). d = –0.5, on the other hand, would represent a moderate negative effect on whatever you decided to measure. Effect sizes, although useful in helping to interpret your findings, do not tell you if you have actually found something, or not, with any degree of certainty. This needs a different approach.

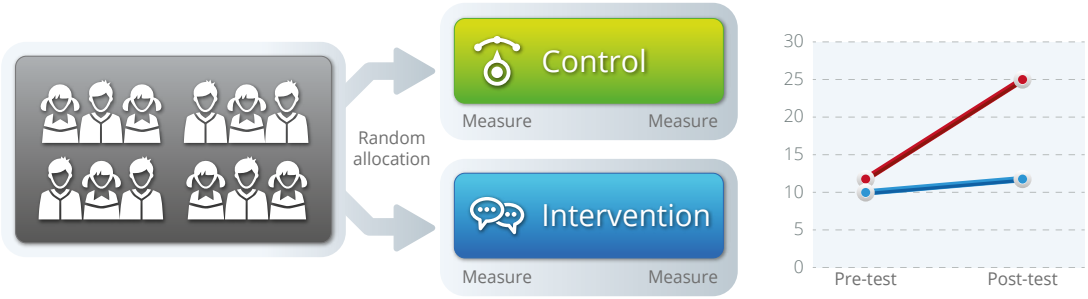
To establish if there is a change in the first place, the convention is to report the level of significance. This works in the following way. Tests called inferential tests are used. There are different tests depending on the type of design you have, the way your data is spread out within your results (the distribution) and the type of data you chose for your test. After doing your research, you apply one of these tests (e.g. the Mann-Whitney U test). All the inferential tests produce a p-value (e.g. p = 0.03). This level tells you the probability that your result may have occurred by chance. p = 0.50 is a fifty in a hundred probability of a chance result (50/50, like tossing a coin), p = 0.01 a one in a hundred probability of a chance result etc. The convention across all scientific fields for a century has been that you only consider that you have found something if you have found a level less than p = 0.05 (or better). Within the formulae for these tests, the statistics work by combining the strength of the effect with your sample size. Hence, a small-scale study might find a large effect that was not significant but a large-scale study could find that a small effect is significant. It is also possible to decide to be more stringent and set yourself a more challenging level (which you might do with an invasive procedure (e.g. p = 0.01 or 0.001)), in which case you would need to build a larger sample size. This said, you never set the threshold level for significance less stringent than p = 0.05.

2. Between-subject designs

2.1 | What is a between-subject design and when can it be used?

Between-subject ('between participant' or 'independent measures') designs are the types of experimental research with which most people will be familiar. In between-subject designs different participants experience the control condition compared to the intervention (also known as the experimental condition) (see Figure 1).

Figure 1: A between-subject design with a pre- and post-test



Between-subject designs have advantages that include being relatively easy to set up (compared to some of the other designs we will look at later). In addition, these forms of research design are appropriate for a wider range of situations than other approaches. However, like all types of design they have a number of disadvantages. Most notably, there is the obvious limitation that you are comparing the responses of different people. Therefore, there is always a risk that variation between the participants caused your results and not the intervention. The risk of this type of error reduces if one uses more sophisticated randomisation. One such approach is stratified randomisation, in which the researcher ensures an equal balance of participant characteristics (e.g. gender, SEN and prior attainment). Deploying a pre-test as well as a post-test (where possible) can help by giving a degree of assurance that at least, with regard to the test you used, people started from the same place – although this may not always be the case (something that you have to take into account in your interpretation of the findings).

Another common way to deal with the naturally occurring differences between the participants who did the control and the intervention, in between-subject designs, is to have a much larger sample size. The assumption here is that as the number of participants increases, so the sample is more likely to approximate to the actual population reducing the potential for error variance. It is also common, in psychology, to mitigate the risk by having a very defined and similar group of individuals taking part in the study right from the start. All of this said, the risk of your research being confounded by participant variation is never fully removed. At the end of the day, there is no perfect design and there will always be limitations with the ultimate answer to this problem, from the perspective of scientific method, being to replicate the research in a number of different contexts perhaps with refinement learning from early iterations. On this point perhaps rests the most exciting future for teacher-led randomised controlled trials – the ability to develop scale and multiple replications whilst engaging teachers more critically in the question of what is evidence and how you can develop it.

In Examples 1 to 3, which follow, you can see a number of teacher-led pieces of experimental research with variations in whether the designs used a pre- and post-test or just a post-test and whether they applied some form of stratification to the randomisation or not. You may also want to note the following other interesting features of their research and how the teachers dealt with various practical as well as interpretive questions.

James Siddle's research (Example 1) is of particular interest because he has built sample size by adopting what is, in other fields, known as a 'Team Science' approach. Team Science involves a single disciplinary, or multiple disciplinary, collaboration within a single trial in order to share resources and build sample size across a number of different trial sites (which can sometimes be international in their location). By pooling resources across 10 rural English primary schools, James has built a sample size able to detect a significant effect even in his smaller subgroups (such as children with special educational needs). His research also illustrates the potential of adopting the style of study more often found in laboratory-style psychology research which takes place over a short timescale (in this case a week). By doing this he has arguably avoided the problem which many six-month-long education trials seem to be facing, i.e. the smoothing out or diffusion of results because so many extraneous variables are likely to confound the research over such a long timescale – a likely outcome facilitated by inherently volatile and unstable social contexts, such as schools.

Example 2, a study by Theresa Peacock and Bridie Bear, also illustrates the fact that even small sample sizes can produce significant results, providing there is a good basis for the protocol being tested and the design is delivered in a tight and consistent way over not too long a timescale. Deploying a robust standardised test arguably also helped. The final between-subject example (Example 3, by Alison Turner, Dean Flood and Kate Andrews, three teachers from Wales) illustrates two further important aspects of classroom-based experimental research – the question of picking an appropriate control condition and how you need to interpret non-significant results in the light of this. Clearly, in most situations doing nothing with a group of children would not be an appropriate control, for it would be beyond doubt that any teacher would be likely to be better than no teaching. Rather, as in the best drug trials, the appropriate control (in most cases) should be the current existing best practice in the school, with the new treatment compared to that best practice. From such a perspective, and this is the case in Example 3, finding a non-significant result (i.e. no difference between the control and the intervention groups) means you have identified an alternative treatment that is as safe, and as effective, as current best practice. Arguably, such results expand the range of teacher professionalism and could offer more cost-effective options – something to be as much welcomed as where treatments with positive significant effects are found.

A final note, relevant throughout this report, is that all the trials adopt the use of the correct effect size depending on whether you have normally distributed data or not, something which has emerged in recent years as a serious issue with the often blanket use of Cohen's d which is the wrong effect size for non-normal data (which should be r, and which in turn needs to be interpreted differently).[3]

[3] Where d = 0.5 is considered a medium (or moderate) effect size, the equivalent threshold for r is 0.3 etc. Other effects sizes, such as np2 (used, for example, to describe the effect when ANOVA has been used) also have different thresholds for interpretation.

1

St Margaret's C E School
Kyra Teaching School Alliance

Verbal and visual-digital feedback on creative writing in rural primary schools improves progress rates compared to written feedback – a preliminary study

Verbal and visual-digital feedback on creative writing in rural primary schools improves progress rates compared to written feedback – a preliminary study

Author:
James Siddle
James.Siddle@st-margarets-pri.lincs.sch.uk

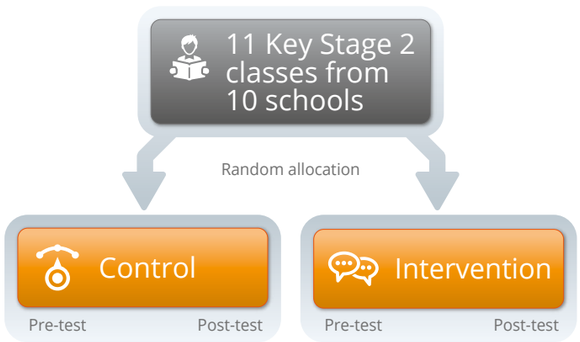
Kyra Teaching
School Alliance

Purpose of the research: Research evidence suggests that effective feedback has a significant impact on pupil progress. Initial trials show the positive impact of digital feedback on outcomes in writing, and the impact may be greatest on SEND (Special Educational Needs and Disability) and FSM (Free School Meals) children. This is an important area to explore using a randomised controlled trial design because it is an approach that is poorly studied at a time when many schools are investing significantly in new digital technology. The study was conducted with the support of a grant from the National College for Teaching and Leadership as part of the Closing the Gap: Test and Learn programme.

The research design

A between-subject design was used with a pre- and post-test. To address the aims of the research the independent variable was operationalised by creating two conditions:

- IV Level 1 (Control condition)
  – Written feedback, the school's normal practice
- IV Level 2 (Intervention) – Digital feedback



Methods

Participants, sample size and randomisation

Eleven classes from ten rural primary schools participated in the study. Pupils were randomly allocated to a control or intervention group in each class. In total, 231 Key Stage 2 pupils (120 boys and 111 girls) took part in the research (113 in control and 118 in the intervention). The total number of FSM pupils was 42 (18.18%), which is below the national average (NA) of 26.6%. The total number of SEND pupils was 40 pupils (17.3%) which is slightly above NA of 16.6%.

Procedures

The randomly allocated groups were given a writing prompt, success criteria rubric and video, together with a short film as a writing stimulus. Pupils had ten minutes' planning and 40 minutes' writing time. The control group received written feedback; the intervention group received feedback digitally. Each group had the same amount of 'fix it' time the following day. Pupils made corrections and recorded 'What I have learnt' statements. Pupils were then given another piece of creative writing (of the same genre) the following day. The procedure was repeated. The work was marked against the two success criteria points and the gain scores were recorded. Blinded marking of approximately 10% of the work was then undertaken.

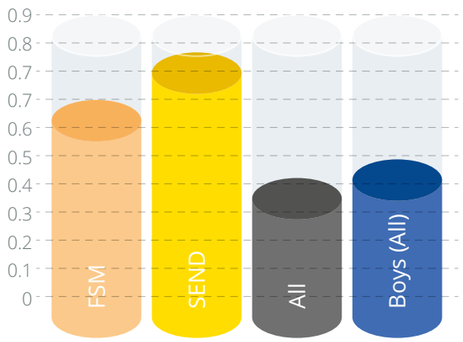
Materials (and apparatus)

A success criteria rubric was used along with a model text. Models of written and digital feedback (through video) were used to standardise marking. A format was given to pupils regarding how to correct their work following feedback.

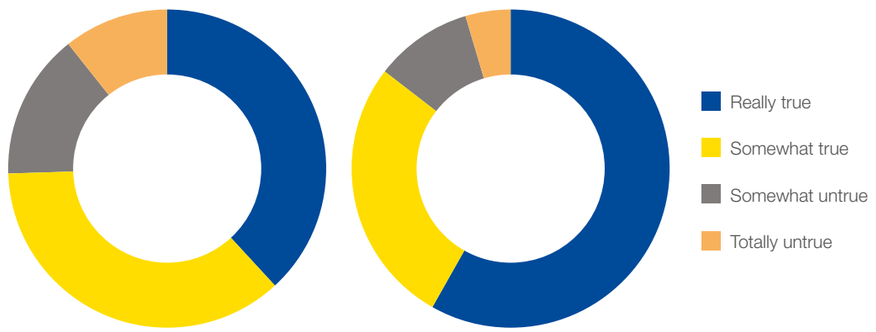
Results

Gain scores were first calculated. Mann-Whitney U tests indicated a significant improvement for all pupils who underwent the intervention compared to the control, and for sub-groups. There was a moderate positive effect size for disadvantaged pupils (n = 43, p = 0.03 (one-tailed), r = 0.308) and SEND pupils (n = 40, p = 0.013 (one-tailed), r = 0.37); and an overall small positive effect for all pupils (n = 231, p = 0.004 (one-tailed), r = 0.218).

Gain scores for all pupils and sub-groups



Pupil perceptions of written vs digital feedback (n = 153)



I think my work gets better following written feedback from my teacher

I think my work gets better following digital feedback via an iPad from my teacher

Limitations

The trial was limited by its relatively small sample size and therefore requires replication with greater numbers. Although the results suggest a greater impact on boys it is not clear why this is the case. Although the effect of the intervention was greatest on SEND pupils the trial did not take into account the specific different needs of these pupils.

Conclusions and recommendations for future research

The gains in the present study were similar to prior EEF research evidence, with regard to the impact of digital technology on closing the gap in attainment (which suggested that digital technology may produce gains of +4 months' progress over an academic year). In particular, the data suggested that the intervention produces the greatest gains for disadvantaged and SEND pupils. The survey that looked at pupil perceptions indicated that, in general, pupils feel they make better progress following digital feedback, evidence which backs up the findings in the RCT. Previous research has also suggested that gains may be even more substantial in mathematics; therefore a future study may wish to look at different subject areas. A future study may also wish to take into account different types of SEND pupils and any difference in effect depending on type of special need.

Kyra Teaching School Alliance is part of CfBT Schools Trust

2

Monks Abbey Primary School

Peer reading improves the reading age of pupil premium children compared to reading only to adults – a preliminary study

Peer reading improves the reading age of pupil premium children compared to reading only to adults – a preliminary study

Authors:
Theresa Peacock and Bridie Bear
Theresa.Peacock@monksabbey.lincs.sch.uk

Purpose of the research: Recent research carried out by teaching schools on behalf of the National College for Teaching and Leadership's (NCTL) national agenda R&D project suggests that peer reading can have a positive impact on pupils' reading ability and enjoyment. The original research shows that it has a positive impact on KS3 pupil premium (PP) pupils and we wanted to explore the potential impact on KS1 and KS2 PP pupils. This was an important area to explore using a randomised controlled trial design because all schools strive to close the gap for all groups. It may also lead to ways to optimise learning time for PP pupils. The research had two aims; these were:

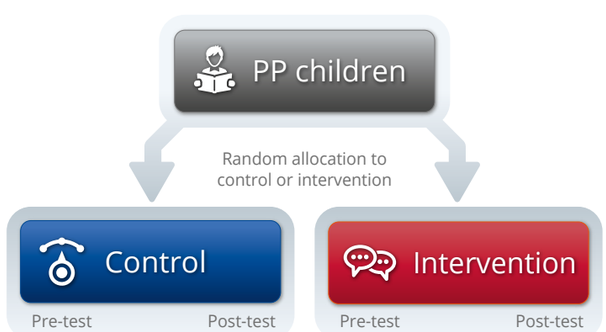
- To establish whether reading to peers can have a positive impact on the reading ability of PP pupils
- To establish whether reading to peers can have a positive impact on the reading enjoyment of PP pupils

This study was conducted with funding from the NCTL Closing the Gap: Test and Learn programme and support from CfBT Education Trust.

The research design

A between-subject design was used with a pre- and a post-test. To address the aims of the research the independent variable (peer reading) was operationalised by creating two conditions:

- **IV Level 1 (Control condition)** – PP readers continue to receive current reading intervention
- **IV Level 2 (Intervention)** – PP readers receive additional peer reading time three times a week



Methods

Participants, sample size and randomisation

Eight classes from an inner-city primary school participated in the study. From these classes, PP children were identified and then randomly allocated to a control or intervention group in each class. As this participant group contained similar pupils and the study primarily aimed to test the effectiveness of the design, simple randomisation was applied.

In total, 54 PP pupils took part in the study. The small sample made it unlikely that anything other than a large effect size would be detected as significant; however, it was considered important to establish the effectiveness of the design before considering the implementation of a larger study.

Procedures

The randomly allocated groups were both given the New Group Reading Test (NGRT) to establish reading age. They were also asked to rank their enjoyment of reading on a scale from 1 to 10.

The control group then continued with the normal reading provision (guided reading once a week, reading to an adult individually once a fortnight).

Members of the intervention group were buddied up with a peer from their own class (working above the reading level of the intervention participant). The intervention group had three 15-minute sessions where they would read to their buddy. This was repeated each week for six weeks. At the end of the six weeks, both groups were again given the NGRT reading age test.

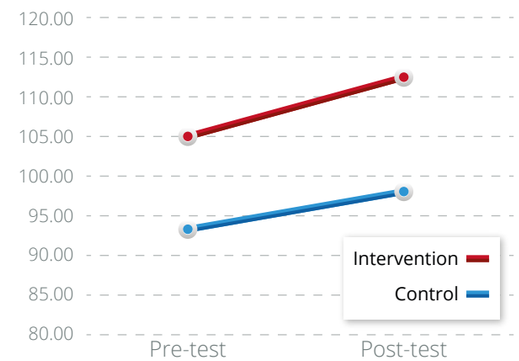
Materials (and apparatus)

The reading buddies had reading records that they would fill in for each other. The NGRT was used to gain a reading age in months. This NGRT was developed by GL Assessment and the National Foundation for Educational Research (NFER) and is available from www.gl.assessment.co.uk. The NGRT includes sentence and passage comprehension.

Results

Gain scores were first calculated using the results in the graph below. A Mann-Whitney U test indicated that there was no difference ($p = 0.114$ (one-tailed)) between the progress rate of children in the intervention (Mdn gain = 5.0) compared to the control (Mdn gain = 2.00). The effect size was small ($r = 0.124$). However, artificially amplifying the sample by a factor of two (from $n = 54$ to $n = 108$) yielded a significant result ($p = 0.043$), suggesting that a future larger study might be able to detect a positive benefit for peer reading compared to existing practice.

Pre- and post-test scores for the control group and intervention



Limitations

The main limitation was sample size. However, it should also be acknowledged that the use of simple randomisation may have introduced the risk of between-participant variation which could have affected the results.

Conclusions and recommendations for future research

The research design was effective in producing findings that suggested that the intervention group made an average of five months' reading age gain over six weeks compared to two months for the control group, although a large study would be needed to confirm this effect. On the current evidence, the intervention appears to be at least equal to existing practice and therefore a viable alternative treatment, one which might show a modest benefit if the findings were replicated in a larger trial involving at least twice as many children. A future study may also wish to consider case-matching or stratified randomisation as a means of controlling for between-pupil variation.

Kyra Teaching School Alliance is part of CfBT Schools Trust

3

Kings Monkton School

Preliminary evidence from a small-scale randomised controlled trial into the effectiveness of a ‘RUCSAC’ individual checklist approach

Preliminary evidence from a small-scale randomised controlled trial into the effectiveness of a 'RUCSAC' individual checklist approach

Authors:

Alison Turner, Dean Flood and Kate Andrews

alisonturner@kingsmonkton.org.uk

KINGS MONKTON SCHOOL

Purpose of the research

Numerical reasoning is an area of high priority for both Kings Monkton School and Mary Immaculate RC Primary School. Data analysis has shown that pupils are performing at higher levels within procedural mathematics as opposed to numerical reasoning. A priority for both schools is to raise and match attainment within numerical reasoning and to improve pupils' independence in applying their procedural skills in real-life tasks. 'RUCSAC' provides a strategy for working through numerical reasoning tasks. RUCSAC is an acronym for 'Read, Underline (Understand), Calculate (Choose), Solve, Answer and Check'.

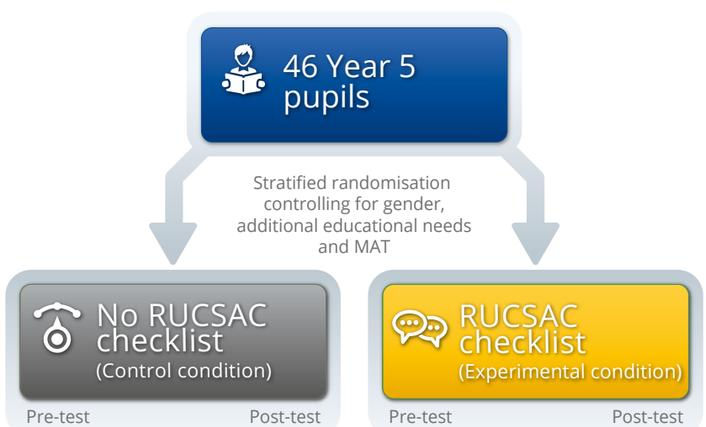
Both schools aimed to test the effectiveness of pupils' use of an individual 'RUCSAC checklist' whilst working through reasoning tasks.

This study was conducted with the assistance of the National Support Programme in Wales and CfBT Education Trust's Experimental Research Design for Teachers two-day training programme.

The research design

A between-subject design was used with a pre- and post-test of numerical reasoning, together with one Likert question with a seven-point scale (assessing confidence). The independent variable (a 'RUCSAC checklist') was defined operationally by creating two conditions:

- IV Level I (Control condition) – No RUCSAC checklist
- IV Level II (Experimental condition) – Use of the RUCSAC checklist



Methods

Participants, sample size and randomisation

Two primary schools, one independent and one maintained, each with single-form entry, took part in the research. All participants were Year 5 pupils, with a total sample size of 46. This enabled the stratified random allocation of individual pupils to control or intervention, controlling for gender, additional educational needs and MAT (more able and talented) pupils.

Procedures

The RUCSAC approach is a systematic strategy which provides an approach that pupils can follow when tackling numerical reasoning tasks. In the opening part of the lesson, different teachers were used for the control classes and the intervention classes to avoid potential contamination between the groups.

The RUCSAC approach was shared with the whole class, with a wall display showing RUCSAC that pupils could refer to throughout their activities. The pupils in the intervention group were given a RUCSAC checklist and asked to check against the success criteria throughout the lesson. The checklist was stuck into pupils' books alongside their calculations.

A number of numerical reasoning lessons were taught to the class over a five-week period.

Materials (and apparatus)

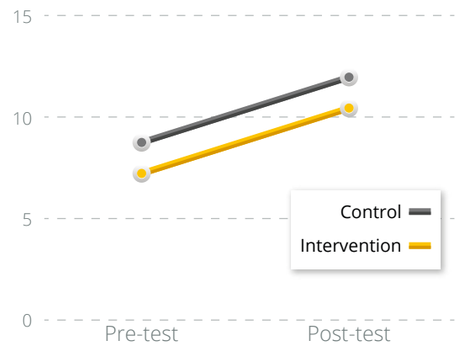
The RUCSAC checklist was used by each pupil in the intervention group. Various numerical reasoning (NR) activities were used by teachers to develop reasoning skills. The May 2014 NR test was used for the pre-test and the May 2015 NR test was used for the post-test. Raw scores results were used. Pupils also rated how confident they felt when working through numerical reasoning activities on a seven-point Likert scale. The research was conducted using the approaches and research methods materials that were trained on the CfBT Experimental Research Design for Teachers programme as part of the National Support Programme in Wales.

Results

Gains scores were analysed. A Mann-Whitney U test (one-tailed) indicated no significant ($p = 0.397$ (one-tailed)) improvement in progress for the pupils who were exposed to the RUCSAC checklist. A very small effect size ($r = 0.045$).

Data from the pupils' confidence questionnaire was not normally distributed, therefore a Mann-Whitney U test was also applied. This showed that there was a significant ($p = 0.013$ (one-tailed)) improvement in pupils' confidence when answering numerical reasoning questions using the RUCSAC checklist. A large effect ($r = 0.605$).

Mean results for the pre- and post-tests for the control and intervention groups



Limitations

There was a small sample size of 46 pupils, therefore the results should be interpreted with caution. In addition, only half the pupils completed the confidence questionnaire. The research was carried out in two separate schools and therefore numerical reasoning activities throughout the five-week period varied to some degree.

Conclusions and recommendations for future research

The progress of learners in numerical reasoning did not appear to be significantly affected by the use of the RUCSAC checklist. Therefore checklists were shown to be a strategy equal in value to existing practice in terms of attainment. However, evidence from the Likert scale questionnaire showed that the checklist did impact positively on pupils' confidence with numerical reasoning. This suggests that checklists might be positive tools in developing pupils' independent learning skills.

In some cases, teachers try to improve pupils' confidence by reducing the level of challenge within a lesson. Because checklists appear to improve confidence, whilst maintaining attainment levels, this suggests that the approach could be a useful alternative strategy for teachers to use. Future research (within a larger replication) may want to look at the use of checklists outside of numerical reasoning (e.g. with grammar or punctuation).

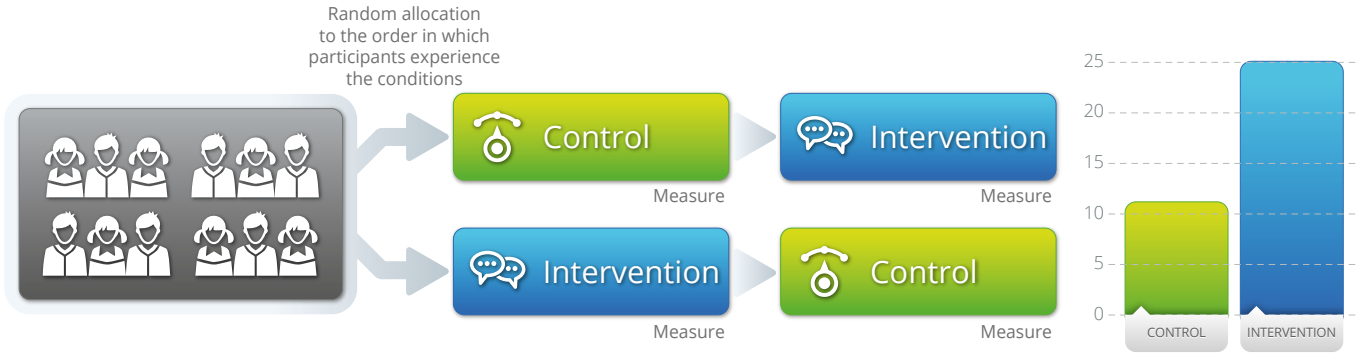
In partnership with CfBT Education Trust

3. Within-subject designs

3.1 | The advantages and use of a within-subject design

Within-subject (or 'repeated-measures') designs seek to resolve the main disadvantage of the between-subject design, i.e. the risk that your results may actually represent naturally occurring differences between the participants rather than the difference between the intervention and the control. Within-subject designs achieve this by getting all the participants to complete all of the conditions within the study (for example, in a two-condition design, with a control and an intervention, the participants would be exposed to both the control and the intervention). By doing so, each participant effectively becomes their own baseline control (and in a way their own pre-test) with the difference that is being measured occurring 'within' the participants (hence the term 'within-subject') rather than 'between' them (Figure 2). This said, it can be desirable to have a pre- and post-test within each condition depending on the research question being addressed.

Figure 2: A post-test only within-subject design (with counterbalanced conditions)



Because between-participant variation has been reduced to a minimum within such designs, the statistics you are allowed to use to analyse a within-subject design are also more forgiving. You therefore can detect a significant effect with a much smaller sample size than in a between-subject design – making such designs ideal for small-scale classroom-based research. In addition, if there were two conditions, you would effectively be counting every participant twice and therefore get a much more efficient use of participants and greater statistical power. For example, combining the two benefits above, a 156-participant two-condition between-subject study (with normally distributed data),[4] would be capable of detecting a 0.4 effect size as significant. However, to detect the same effect size in a within-subject design would only require 41 people to take part – the result of being able to apply a paired-sample t-test rather than an independent sample t-test and the double counting described above.[5]

At this point, you might well be thinking 'Why do experimental research designs and RCTs not use a within-subject design as a matter of course?' As with all types of research design, there are advantages and disadvantages, and remaining limitations even when you have completed your research. The main disadvantage in a within-subject design is that people tend to get better when they do something twice. Therefore, they may end up responding to the second condition that they experienced in a different way because of what happened to them in the first condition. When this occurs, we call it a 'carry-over' effect (or 'order effect'). For this reason, within-subject designs

[4] 78 in the control and 78 in the intervention.
[5] Calculated using G*Power 3.1.7, one-tailed with alpha = 0.05 and 80% power (see www.gpower.hhu.de, for references, registration and software download).

usually include 'counterbalancing'. This involves randomly allocating the participants to two different condition orders with the conditions reversed (Figure 2). The researcher may also feel it necessary to leave a time gap between the two conditions to allow the effects of the previous one to wear off; the idea with counterbalancing being that if there is, for example, an inflation in the results because an effect has carried over from the first to the second condition, then the same inflation will be present in both the intervention and the control in equal amounts since they both occupied the second condition space. In which case, the effects theoretically should cancel each other out.

There is a final more important limitation with a within-subject design. You cannot conduct a within-subject design if the effects of your intervention are irreversible (i.e. if you cannot remove the effects of the intervention once experienced). For example, in a classroom, although it is possible to have a within-subject design where you are testing different pedagogical processes (group work versus individual work with very similar topics as the lesson content), it would not be possible if the intervention was content-bound (for example, learning new mathematics approaches versus not learning them and then doing a mathematics test). In such a case, you need to adopt a between-subject design (or a matched-pair design (another form of design we will discuss later)).

The next two examples follow exactly the process described above. The first two (Example 4, by Daniel Lear and Example 5, by Matthew Maughan and David Ashton) fascinatingly, were both the result of attending the same Experimental Research Design training (delivered in Wales between March and July 2015). The two research teams did not work together during the design day, but at the analysis and interpretation day found that they had both identified the same research question and had both found significant benefits for the use of flipped learning[6] with their mathematics students. This is an exciting finding because it represents the first example of replication within the newly emerging field of teacher-led experimental research. The first of the final two within-subject design examples (Example 6, by Timm Barnard-Dadds and Allison Davies) demonstrates the value of finding a likely alternative treatment with implications for teacher practice. Example 7, by Sarah Baugh-Williams, Ceri Bibby and Graeme Jones shows how a small-scale trial can lay the groundwork for a future larger replication by demonstrating the effectiveness of the design – despite not finding a significant result because of the sample size.

[6] Content delivered through online video prior to the lesson

4

Cwmtawe Community School – Ysgol Gymunedol Cwmtawe

Two mathematics lessons of flipped learning improve performance in numerical reasoning tasks for Key Stage 3 students

YSGOL GYMUNEDOL CWMTAWE COMMUNITY SCHOOL

Two mathematics lessons of flipped learning improve performance in numerical reasoning tasks for Key Stage 3 students

Author:
Daniel Lear, Cwmtawe Community School
dan_lear83@hotmail.com

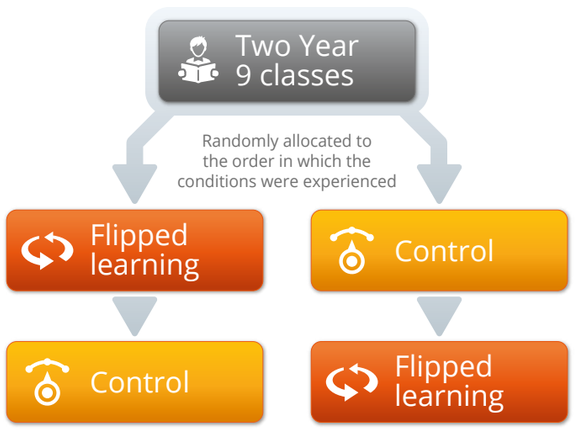
Purpose of the research: With the introduction of a Numeracy GCSE (from 2015) being a follow-on from numerical reasoning at Key Stage 3, Cwmtawe Community School is gaining an extra hour's teaching time per fortnight at Key Stage 4. With the use of a flipped learning environment in this extra hour, it is hoped pupils will take greater responsibility for their own learning, ultimately leading to improved performance in the new Numeracy GCSE. This trial considered whether flipped learning (content delivered through online video prior to the lesson) and the extra contact time in class would benefit pupils and as a result raise attainment in numerical reasoning style questions. The research was conducted with the assistance of the National Support Programme in Wales and CfBT Education Trust's Experimental Research Design for Teachers two-day training programme.

The research design

A within-subject design was used with a pre- and post-test. To address the purpose of this research, the independent variable (flipped learning environment) was operationally defined by creating two counterbalanced conditions:

IV Level I (Control condition) – Normal classroom practice

IV Level II (Experimental condition) – Pupils receive content teaching through an online video prior to the lesson



Methods

Participants, sample size and randomisation

Two Year 9 mathematics classes at Cwmtawe Community School (pupils achieving Level 5 at Key Stage 3) took part in the trial, consisting of 17 girls and 23 boys. Pupils were all aged 13–14. These pupils were selected as they will become the first year group to follow the new Numeracy GCSE syllabus. The classes were randomly allocated to the order in which they would experience the conditions.

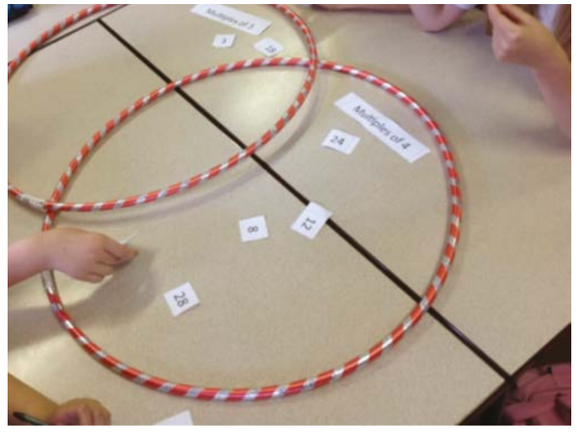
Procedures

The trial consisted of two lessons with lesson plans remaining constant for both intervention and control groups. Lessons were taught by the same teacher, who maintained consistent practices throughout. Group 1 (after random allocation) participated in flipped learning for lesson 1 and normal classroom practice for lesson 2. Group 2 followed normal classroom practice for lesson 1 and flipped learning for lesson 2.

Prior to the trial, pupils sat a numerical reasoning test (pre-test). For each lesson, there was a multi-stage question of increasing difficulty in the test. At the conclusion of each lesson the pupils again sat the test question pertinent to that lesson (post-test).

Materials (and apparatus)

The test was devised in-house based on sample materials from the new Numeracy GCSE and LNF numerical reasoning papers. The flipped learning materials were made available to pupils on the school website and via YouTube.

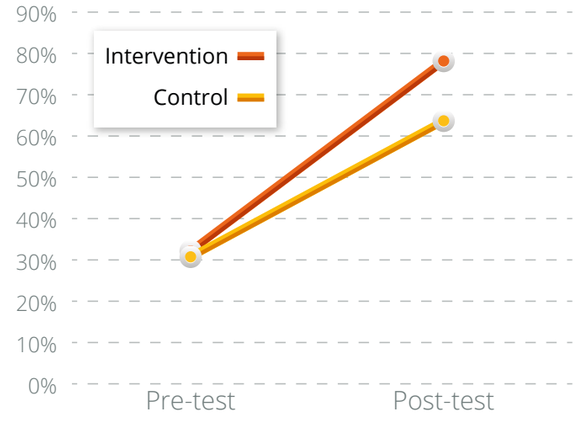


Results

Gain scores (calculated from pre- and post-test scores) were used in the analysis. A Wilcoxon signed-rank test indicated a significant ($p = 0.001$ (one-tailed)) improvement in progress for the pupils who were exposed to the flipped classroom method (Mdn = 0.50) compared to the control (Mdn = 0.38). A large effect size ($r = 0.459$) was noted.

Included in the study were Year 9 pupils currently achieving Level 5 at Key Stage 3. A replication would be required to ensure the conclusion holds true for more able and talented or lower-ability pupils.

Mean pre- and post-test scores for intervention (flipped learning) and control groups



Conclusions and recommendations for future research

Flipped learning improves attainment in numerical reasoning. Teacher observations during the trial suggest this result may have arisen from benefits such as extra time for pupil–pupil and pupil–teacher interaction within the classroom and freedom to watch the content at a time, place and pace suitable for the individual. It is not clear from the results whether the improvement is a direct result of one of these facets or a culmination of all. It was good to see pupils spending lesson time actively doing maths rather than passively watching the teacher do maths on the whiteboard. Although there was a small sample size in this research, a parallel study (Bassaleg School, Newport, included in this journal) indicates a similar result for the flipped classroom.

In partnership with CfBT Education Trust

5

Bassaleg School – Ysgol Bassaleg

The use of flipped learning, prior to beginning a new concept in mathematics, has a positive effect on pupils' learning

The use of flipped learning, prior to beginning a new concept in mathematics, has a positive effect on pupils' learning

Authors:

Matthew Maughan
matthew.maughan@newport.gov.uk

David Ashton
mrashtonmaths@gmail.com

Purpose of the research

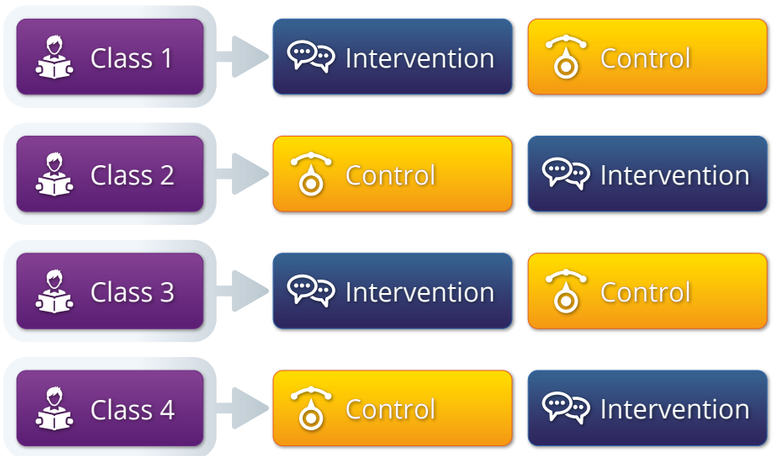
Bassaleg School uses flipped learning as part of a digital literacy drive. The school has recently removed textbooks from mathematics classrooms in the belief that repetition of simple problems in class reduces the time spent developing a detailed understanding of the concept in question, and that flipped learning techniques are a more efficient way of establishing understanding. There is only anecdotal evidence to date that this technique has a positive impact.

This study was conducted with the help of the National Support Programme in Wales and CfBT Education Trust's research methods training programme.

The research design

A post-test only, within-subject design was used. To address the aims of the research, the independent variable (the use of flipped learning) was operationally defined by creating two counterbalanced conditions:

- IV Level I – Control condition: experience of new material in classroom as per normal practice
- IV Level II – Experimental condition: use of flipped learning prior to the lesson



Methods

Participants, sample size and randomisation

The participants in the research were Year 7 students in mathematics classes. Because the classes were in bands of ability set, four similar classes participated. Stratified randomisation ensured that each class had an equal mix of male and female students and a similar spread of ability.

The testing procedure had four slots into which a class could be randomly placed, two experiencing the flipped learning independent variable first and a control condition second. Two classes counterbalanced this and experienced a control first and the independent variable second.

Procedures

Two concepts were selected, of similar challenge, which pupils had not experienced before. These were: enlargements and rotations.

The first series saw two of the four classes follow a flipped learning programme prior to the lesson; the flipped learning was shared so that staff could monitor whether pupils accessed the material appropriately. All four classes then experienced the same lesson including a post-intervention test to assess pupil understanding.

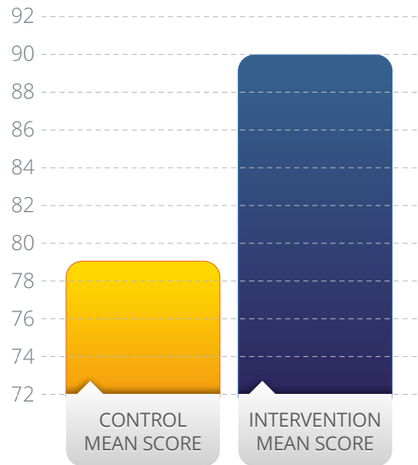
The same process was then followed for the counterbalance testing, with the pupils exposed to flipped learning from the first sample becoming control pupils for the second.

Materials (and apparatus)

Identical lessons, planned by the same teacher, were shared with the teaching team. Pupils exposed to the intervention were given a search engine link containing flipped learning material. Staff monitored this to ensure the material was accessed by the sample group. The test used consisted of unseen questions, from a worksheet, appropriate to each of the topic areas introduced.

Results

A Wilcoxon signed-rank test indicates a significant (p=0.002 (one-tailed)) improvement in attainment for pupils who were exposed to the flipped learning method (Mdn = 100), compared to the control (Mdn = 90) – a moderate effect size (r = 0.26).



Mean control scores and mean intervention scores (flipped learning)

Limitations

The study was limited to students working at the upper end of expected outcomes in mathematics for their year group (Level 5, Year 7).

Conclusions and recommendations for future research

The research suggests that pupils' experiencing flipped learning techniques before they embark on unfamiliar topics within mathematics has a significantly positive effect on attainment (p = 0.002). Although the sample size was relatively small, a parallel study conducted simultaneously (at Cwmtawe Community School, Swansea, also reported in this journal) yielded similar results. Future research may want to seek to replicate the findings with different groups of children.

In partnership with CfBT Education Trust

6

Pentre’r Graig Primary School
Cwmrhydyceirw Primary School

A preliminary pilot study into the effectiveness of a ‘rich task’ contextual style of teaching mathematics, compared to a traditional procedural approach

"Learning, Caring, Having fun!"
"Dysgu, Caru, yn cael Hwyl!"

A preliminary pilot study into the effectiveness of a 'rich task' contextual style of teaching mathematics, compared to a traditional procedural approach

Authors:
Timm Barnard-Dadds and Allison Davies
cwmrhydyceirw.primaryschool@swansea-edunet.gov.uk
pentrergraig@swansea-edunet.gov.uk

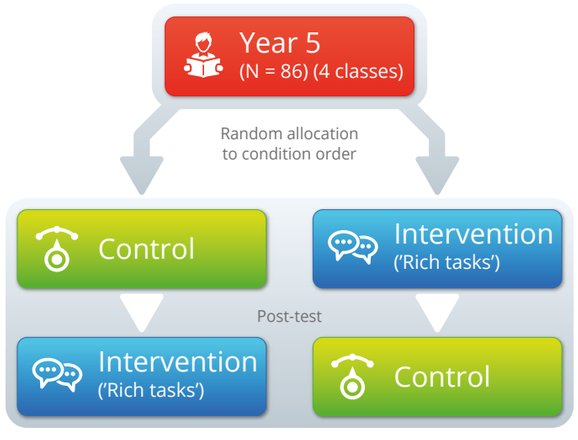
Purpose of the research: Educators believe pupils understand mathematical concepts better, and are able to recall them, if the concepts are taught in a meaningful context rather than learnt by rote or simply through repetition of examples (Welsh Government (2013) *National Literacy and Numeracy Framework: to support schools in introducing the National Literacy and Numeracy Framework*). 'Rich task' contextual teaching involves teaching a skill and applying it at a later stage in a meaningful context. Conversely, traditional 'procedural' teaching involves teaching a method that can then be replicated to solve given problems of a similar type. This research aimed to establish whether the approach had either a negative or a positive effect on pupil attainment, or no effect (which would mean that it was an acceptable alternative that 'does no harm'). This study was conducted with the assistance of the National Support Programme in Wales and CfBT Education Trust's Experimental Research Design for Teachers two-day training programme.

The research design

A post-test only, within-subject design was used. To address the aims of the research the independent variable (the use of 'rich-task' contextual teaching) was defined operationally by creating two counterbalanced conditions:

IV Level I (Control condition)
– Procedural teaching in one lesson

IV Level II (Experimental condition)
– Rich task approach in another single lesson



Methods

Participants, sample size and randomisation

Two primary schools in Morriston, Swansea took part in the research. One school has single-age Year 5 classes and the other has mixed Year 5/6 classes. In total the study involved 86 children (41 girls and 45 boys). The pupils were taught in mixed-ability classes randomly allocated to the order in which they experienced counterbalanced conditions.

Procedures

For the first topic, one class of pupils experienced a lesson using a traditional procedural approach to teaching and the other class had a lesson delivered using a rich task contextual approach. For the second topic, the first class had a lesson delivered using a rich task contextual approach and the second class experienced their lesson using the procedural-style approach. At the end of each lesson a test of mathematical knowledge and understanding was administered. This test used the style of questions from the Welsh Government National Numeracy Procedural Test.

Materials (and apparatus)

Materials for teaching a mathematical concept using a rich task contextual approach and a traditional procedural approach were developed by the class teachers. These were delivered to pupils prior to the administration of the test questions.

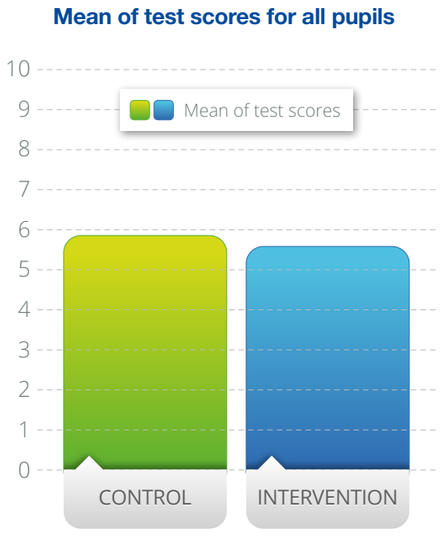
Results

A Wilcoxon signed-rank test indicated a non-significant (p = 0.329 (one-tailed)) improvement in attainment for pupils who were exposed to the rich task contextual lesson (Mdn = 5.00) compared to the traditional procedural control lesson (Mdn = 6.00). A small negative effect was noted (r = –0.043).

Because the data showed that attainment was equal for all learners in both types of lessons, results were then disaggregated to enable the analysis of the attainment of different groups of learners (SEN, average ability and high ability). SEN and average-ability pupils made no significant improvement. However, for high-ability learners, rich task contextual teaching approached significance (r = 0.08).

Limitations

The research was limited due to the sample size involved, and it may also have been affected by the choice of post-test material used to measure attainment, which was unstandardised.



Conclusions and recommendations for future research

This research suggests that using a rich task contextual-style approach to teaching mathematics can be seen as being equivalent in effect to using a traditional procedural style in developing mathematical understanding. In addition to this, it was noted that higher-ability learners may be more able to apply their mathematical knowledge in rich task contextual lessons. A possible reason for this could be that higher-ability learners are more able to apply their knowledge of mathematical concepts as they are already secure in their knowledge of procedural methods. A future study may wish to replicate the design with a sample of high-ability learners in excess of 150–200 in order not to miss a repeated small effect.

In partnership with CfBT Education Trust

7

Y Bont Faen Primary School
Gwaunfarren Primary School
Hafod Primary School

Using a story map approach can be an alternative treatment when solving reasoning problems – evidence from a small-scale preliminary study

Using a story map approach can be an alternative treatment when solving reasoning problems – evidence from a small-scale preliminary study

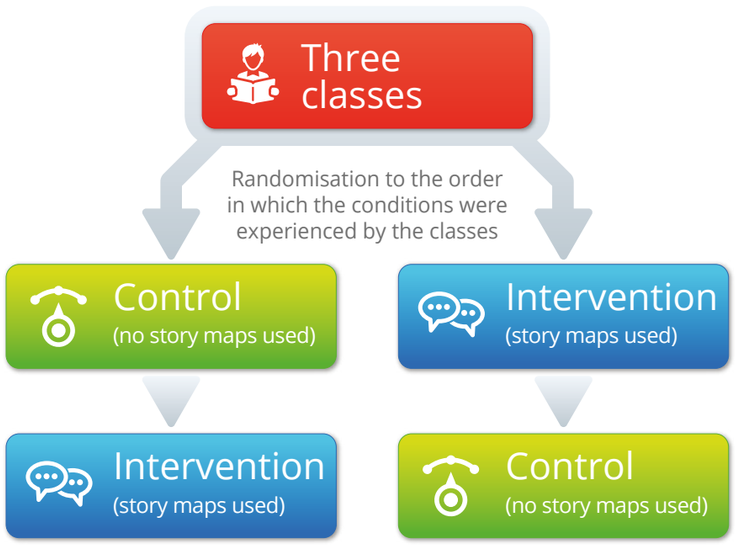
Authors:
Sarah Baugh-Williams, Ceri Bibby, Graeme Jones
sarahbaughwilliams@hotmail.co.uk

Purpose of the research: Across all three schools involved in the research, the improvement of reasoning skills is seen as a key school improvement priority. This study therefore aimed to answer the question: Can the use of pictorial representations in the form of story maps improve the reasoning skills of primary-age pupils? Thus it sought to establish whether this approach had a positive effect on attainment, or no effect (which would mean that it was an acceptable alternative treatment that 'does no harm'). This study was conducted with the assistance of the National Support Programme in Wales and CfBT Education Trust's Experimental Research Design for Teachers two-day training programme.

The research design

A within-subject design was used with two counterbalanced conditions. The independent variable (using story maps) was operationally defined by creating two conditions. A post-test only design was used.

- IV Level I (Control condition) – No story maps
- IV Level II (Experimental control) – Using story maps



Methods

Participants, sample size and randomisation

Three similar-sized primary school classes took part. The classes were from three different primary schools and included two Year 3 classes and one Year 2 class. Children were aged between six and eight years. In total the sample included 81 pupils. All three classes were randomly allocated to the order in which they experienced the conditions within the study. There was a control group and intervention group in each class.

Procedures

Before commencing the research, each teacher delivered a series of lessons to all pupils to show them how to use story maps to support reasoning questions. Teachers jointly agreed and planned which two questions to use for testing.

The teachers delivered the two tests in the same order. The intervention and control groups were separated during the test.

Materials (and apparatus)

The study used sample materials from the Learning Wales website:

Test 1 – 'Aliens' Legs'

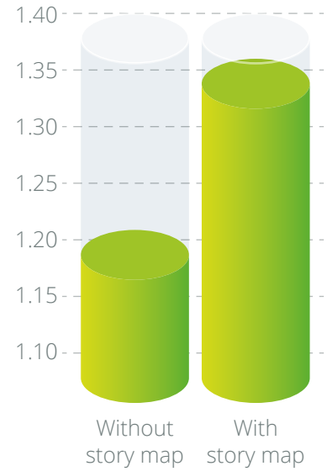
Test 2 – 'Mrs Jones Makes Tea'

The two tests were marked by each teacher using the mark schemes provided to ensure a consistent approach. The first test consisted of a 'real life' situation, in which children were required to calculate the number of cups of tea drunk by a lady throughout the course of a day, using the information provided in the question. In the second test, children were required to draw the appropriate number of legs onto a given number of 'aliens'. They needed to understand and use the information provided in order to show the correct number of legs on each alien.

Results

A Wilcoxon signed-rank test indicated a non-significant effect ($p = 0.081$ (one-tailed)) when the pupils were exposed to the story map method (Mdn = 2) compared to the control (Mdn = 1). A small effect size ($r = 0.08$).

Average scores for reasoning tasks



Limitations

The sample size and number of questions attempted by pupils were too small for the study to make a strong claim regarding its finding.

Conclusions and recommendations for future research

Using story maps had the same effect in this study as not using them when children were solving reasoning problems. However, the small positive effect of story maps approached significance, suggesting the sample size may not have been sufficient. Future research may wish to explore the same treatment in a larger study.

In partnership with CfBT Education Trust

4. Adding a third condition

4.1 | When to consider using a third condition

In some circumstances, the researcher may not consider a two-condition design to be sufficient to control for some things that could confound the research and therefore they may add in an active control. An active control could contain an element of what you are interested in exploring in your experimental condition but not all of the condition – so you can see more clearly which part of the protocol makes a difference and which does not. Another alternative variant could be to test two similar interventions at the same time. The two examples that follow illustrate both of these circumstances.

Example 8 by Gavin Jones and Rob Wilson illustrates the use of an active control. In their study the experimental condition (or intervention) consisted of collaborative learning based around a structured template, their control condition normal practice. To establish the relative effects of the template compared to collaboration, they deployed an active control with just collaboration. Example 9, by Charlotte Morris, shows how you can test two different interventions at once and by doing so make more efficient use of a limited number of participants – an approach which also enables the researcher to interpret the relative effects of the interventions more easily than if the two interventions had been assessed within separate trials. Charlotte's design also incorporated a very innovative use of video to control for teacher variation – by having the instructional elements of the lesson delivered by a single teacher on screen with a facilitator managing the class at various locations. Again, because these designs are within-subject, they have used counterbalancing to control for carry-over effects.

Diagrammatically the two main options of having either a between-subject design or a within-subject design and three conditions are illustrated in Figures 3 and 4. Figure 3 shows a pre-/post-test between-subject design with three conditions.

Figure 3: A between-subject design with three conditions

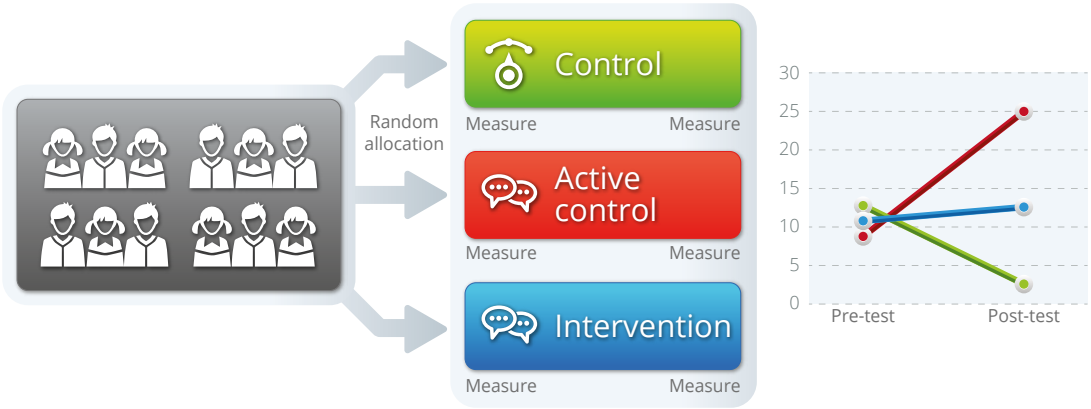
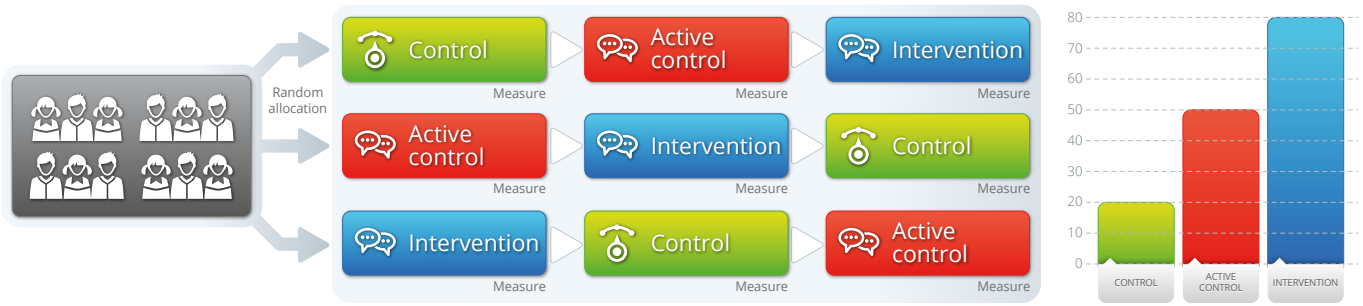


Figure 4 gives the structure for a counterbalanced within-subject design also containing an active control. Notice how all the data eventually will be amalgamated into three columns (control, active control and intervention) – a process that as we discussed before should smooth out carry-over effects because we have counterbalanced.

Figure 4: A within-subject design with three conditions (with counterbalancing)



Having three conditions in your research design adds an additional analytical requirement: namely that the more levels there are to your independent variable (e.g. IV Level I – control, IV Level II – active control, IV Level III – experimental condition) the more likely it is that any change in the dependent variable may have arisen by chance.[7] For example, in a three-condition study the change you detect could have been the result of having tested three things at once and not a true reflection of the difference between any pair of conditions. This problem is known as family-wise error.

The solution to this issue is illustrated in both of the research posters that follow; the convention being that you first conduct some form of ANOVA (analysis of variance) across all three conditions (which essentially tells you if the overall change was significant). Then you follow this with comparisons between each of the pairs of conditions applying a more stringent threshold for significance. This more stringent level is known as a Bonferroni adjustment and involves dividing the threshold you set by the number of tests you will do. The ultimate implication of this is that a larger sample size is usually required for three-condition designs in order to avoid missing an effect.

[7] The independent variable (IV) is what you manipulate as the researcher (i.e. the conditions you allocate people to), the dependent variable (DV) consists of the results which 'depend' on your manipulation of the IV. Typically in education research, the DV is likely to be a test result of some sort, but may also include things such as measurement of pupil perceptions, attendance data, engagement measures etc.

8

Llanishen High School
Hawthorn High School

A collaborative teaching approach enhances the performance of students in mathematical problem solving

A collaborative teaching approach enhances the performance of students in mathematical problem solving

Authors:

Rob Wilson, Llanishen HS, Cardiff
r.wilson@llanishen.cardiff.sch.uk
Gavin Jones, Hawthorn HS, Pontypridd
Gavin.Jones@hawthornhs.co.uk

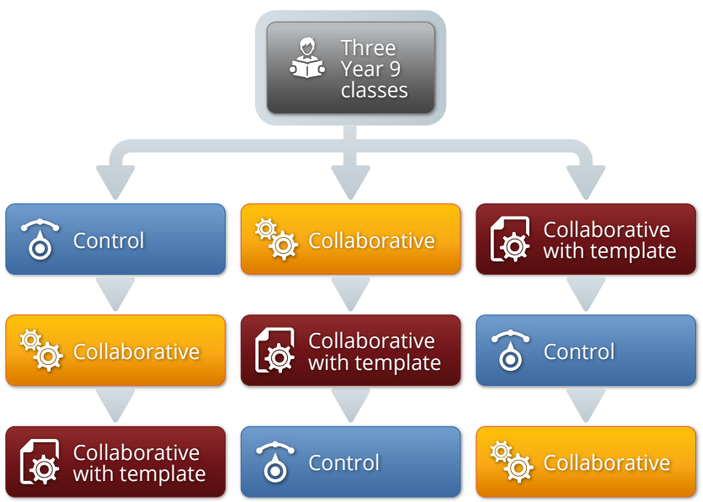
Purpose of the research

The investigation, involving two schools in South Wales, looked at how collaborative learning positively developed student performance and raised confidence levels of students when approaching mathematical problem-solving tasks. Another aim of this research was to establish whether structured collaboration using a template format improved student performance. This study was conducted with the assistance of the National Support Programme in Wales and CfBT Education Trust's Experimental Research Design for Teachers two-day training programme.

The research design

A within-subject design was used with a pre- and post-test. The independent variable (template use during collaboration) was operationally defined by creating three counterbalanced conditions:

- IV Level I (Control condition) – Normal teaching practice: learners are presented with a task and asked to complete it
- IV Level II (Active control) – Collaboration (teacher presents a task; in groups learners discuss possible strategies for completion and then complete the task)
- IV Level III (Experimental condition) – Collaboration and template (teacher presents a task and a template giving a completion structure; in groups the learners discuss possible strategies for completion and then complete the task)



Randomisation to the order in which the conditions were experienced by classes

Methods

Participants, sample size and randomisation

Six Year 9 classes took place in the research, a total of 120 students. A priori power analysis indicated that this sample was sufficient to give 80% power to detect a dz = 0.27 effect size between the conditions, with alpha = 0.017 (Bonferonni adjusted for multiple comparisons (one-tailed)). To attain this level of power with a between-subject design would have required 484 participants, making the use of a within-subject design highly desirable and efficient. The classes, which were predetermined by setting, were then randomly allocated to the order in which they experienced the three conditions using the Excel Rand() function.

Procedures

Students were asked to attempt the tasks using varied pedagogical styles over a series of lessons. Firstly, the control condition involved students being given the problem with basic instructions provided by the teacher and then being given time to tackle the problem. The format changed for the next question (active control), as the students were allowed to discuss possible strategies for completion and then complete the task. Finally, for the experimental condition, students were given a template outlining a suggested structure for completion, then in groups they discussed the possible strategies for completion and then finished the problem set. Teachers had a script to follow, thus avoiding potential biases that might have been caused by a varied approach.

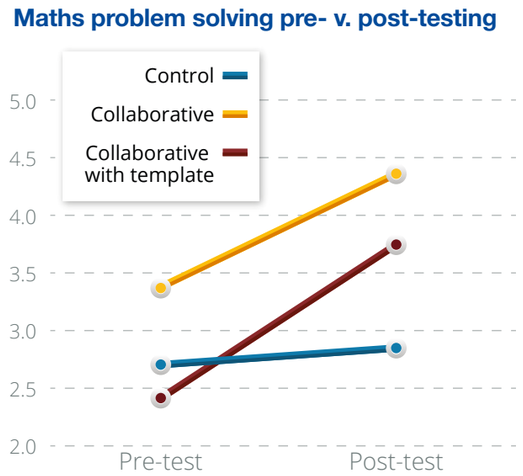
Materials (and apparatus)

The study used the following materials and apparatus. Pupils were given numerical reasoning (NR) 'test style' tasks and, in the intervention group, templates outlining the methodology of task completion. Consistent classroom layouts were arranged between the classes. Test materials consisted of six NR style questions (of equal degree of difficulty) and a seven-point Likert scale confidence measure.

Results

Gain scores were analysed using the pre-post test data in the graph below. A Friedman test showed significant difference across all conditions ($p < 0.001$). Separate Wilcoxon signed-rank tests were used to compare the three conditions to one another. Because the analysis used multiple tests, a more stringent significance level (Bonferonni adjustment) was set (0.017). These results can be found in the table below. This shows that collaborative learning was significantly better than ordinary practice, and collaborative learning with a template was significantly better than both conditions.

A Likert scale confidence questionnaire demonstrated that student confidence levels similarly improved after each treatment.



	Control	Collaborative	Collaborative with template
Control		r = -0.24 p < 0.001 (one-tailed)	r = -0.45 p < 0.001 (one-tailed)
Collaborative	r = 0.24 p < 0.001 (one-tailed)		r = -0.06 p < 0.001 (one-tailed)
Collaborative with template	r = 0.45 p < 0.001 (one-tailed)	r = 0.06 p < 0.001 (one-tailed)	

Limitations

This was a pilot that demonstrated significant results, although further research would be necessary to validate these outcomes. Future research may wish to explore the same treatment in other mathematical situations.

Conclusions and recommendations for future research

Collaborative learning with a template has a large effect (r = 0.45) on mathematics problem-solving and a greater effect than just collaborative learning (r = 0.06). However, collaborative learning alone also makes a moderate difference (r = 0.28). Teachers doing collaborative learning would be advised to combine a template approach with their groupwork activities in order to maximise performance. A possible extension to the trial including the integration of positive and negative visualisation is being considered.

In partnership with

9

The Great Oaks Federation

‘Look, Cover, Check, Write’ improves attainment in Year 1 primary school lessons

‘Look, Cover, Check, Write’ improves attainment in Year 1 primary school lessons

Author:
Charlotte Morris
Charlotte.Morris@gofederation.co.uk

Purpose of the research: This is an important area to explore using a randomised controlled trial design because spelling is a weakness for children throughout the key stages in our school and the children do not always engage in spelling homework. Finding a more active strategy could help create a method that is suitable for all groups of learners. A number of approaches are possible and so the study has aimed to test two strategies against a control condition to ensure efficient use of participants. The research by necessity applied a mixture of one- and two-tailed hypotheses because although it was predicted that both active learning and the Look, Cover, Check, Write strategy would be better than the control, it was not known which of these would be best when compared to each other. This study was conducted with the support of a grant from the National College for Teaching and Leadership as part of the Closing the Gap: Test and Learn programme.

The research design

A post-test (counterbalanced) within-subject design was used. To address the aims of the research the independent variable was operationalised by creating three conditions that allowed for the testing of two interventions simultaneously:

- **IV Level 1 (Control condition)**
  – Normal teacher practice without video delivery
- **IV Level 2 (Intervention A)**
  – Teacher on video delivering general (active) spelling strategy
- **IV Level 3 (Intervention B)**
  – Using the Look, Cover, Check, Write (LCCW) approach delivered by the same teacher on video

Methods

Participants, sample size and randomisation

Three mixed-ability Year 1 classes were randomly allocated to the order in which they experienced the conditions. 88 pupils took part in the study. Prior to analysis, two missing pieces of data (caused by absence) were replaced with the mean for that group.

Materials (and apparatus)

Videos of spelling strategies and a script for the teachers were developed. There were also standard sets of spellings and a test score sheet.

Procedure

Each class had the same set of words to learn during the fortnight. Then each fortnight new words were added to the spelling list. The spelling tests were 10 minutes long and delivered in the mornings. All teachers involved received a detailed briefing prior to the start of the research.

Results

An initial Friedman’s ANOVA indicated that the overall change (shown in the graph) was significant ($p < 0.005$ (two-tailed)) with a moderately small effect size detected ($W = 0.22$). This test was then followed by planned comparisons comparing all conditions with each other using Wilcoxon signed-rank tests. A Bonferroni adjusted threshold for significance of 0.0167 was applied. The results from these tests and effect sizes are given below.

	Control	Active spelling	LCCW
Control		$r = -0.045$ $p = 0.037$ (one-tailed)	$r = -0.241$ $p < 0.005$ (one-tailed)
Active spelling	$r = 0.045$ $p = 0.037$ (one-tailed)		$r = -0.191$ $p < 0.005$ (two-tailed)
LCCW	$r = 0.241$ $p < 0.005$ (one-tailed)	$r = 0.191$ $p < 0.005$ (two-tailed)	

Mean spelling for the three conditions in the present study

Limitations

This study used a more laboratory-style approach than most education experimental research so far. This said, it is believed that the study maintained high levels of mundane realism (maintaining a real classroom environment) and therefore good levels of both external and internal validity. It is too soon, however, to be certain what the effect of these much more tightly controlled forms of design are and whether they produce demand characteristic and other biases (resulting from the use of video rather than a live teacher) that are known in psychology research.

Conclusions and recommendations for future research

Use of the LCCW strategy produced significantly better attainment during the spelling tests than both the active spelling approach and normal classroom practice. However, it appears that the active spelling approach is at least an equal alternative treatment to normal practice. A moderately small positive effect on attainment was detected with regard to the LCCW approach compared to the control and a moderate effect compared to the active learning approach. A future study may wish to look at the effectiveness of the approach in different contexts and with different sub-groups of pupils. In summary, LCCW appears to be a highly effective way of improving children’s spelling as measured by in-class testing.

Kyra Teaching School Alliance is part of CfBT Schools Trust

5. Case-matching and matched-pair designs

5.1 | Another way to deal with between-participant variation

A variant of the between-subject design which attempts to get as close as possible to the main benefit of a within-subject design (the reduction of between-subject variation) is known as a matched pair design. In such a design, participants are first case-matched into pairs. This involves pairing up two people who are very similar to each other according to factors that you think might affect the results (i.e. things that you want to control for). For example, we might identify the two girls with the highest prior attainment who are also of a certain ethnic group, then the next highest-attaining girls etc. Following this, we might then do the same for all the boys. Having case-matched all of our participants into pairs we then randomly allocate each pair to either the control condition or the intervention. The end result of this is two groups of children who are as similar as we can possibly make them – in areas that we think might confound the research if we did not control for them this way. In this way case-matched designs seek to remove the main issue in a between-subject design.

In the examples that follow, you will see a matched pair design with randomisation by Wendy Blyth and Rachel Elphick and a non-randomised case-matched study by Emmet Glackin. The first study (Example 10) illustrates the benefits of using such an approach where you are running a small-scale pilot study, the second (Example 11) how case matching can enhance research in a situation where it is not possible to randomise, and you have to identify a control group to compare your fixed intervention group to.

There is one other advantage to a case-matched study; namely, that by convention, you are allowed to use the same statistics used for within-subject designs – the assumption being that your case matching is appropriate and that you can justify it. In the case of a study with two conditions (e.g. a control and intervention) you can use a paired sample t-test (or Wilcoxon signed-rank test, depending on your data) rather than an independent sample t-test (or Mann-Whitney U test). This means that you end up with more statistical power (the ability to detect an effect) and thus are more likely to be able to detect an effect as being significant.

The relative differences in power for a comparison between two individual conditions at once in a between-subject design compared to a within-subject design/matched pair design are illustrated in Figures 5 and Figure 6, below. Arrows help to illustrate the differences in power between these designs by showing the point at which sample size is able to detect a 0.2 effect size as being significant. Doing so clearly exemplifies the increased efficiency inherent in both within-subject and case-matched studies.

This said, having a sample size of between 300 and 1,000 is always going to be better than running a small-scale study, because you will be able to detect greater levels of significance (if there is an effect there in the first place). As you can see, less than half the number of participants would be required in a within-subject or matched-pair design compared to a between-subject design. These graphs also illustrate a further important point, the fact that you can have too large a sample size and end up wasting resources. This is because, ultimately, effect size detection flat-lines at 0.1 fairly soon, and therefore, you would be better off (and more scientifically minded) if you put your efforts into a replication than carrying out a study with many thousands of participants.

Figure 5: The power to detect an effect between two conditions in a between-subject design with the threshold for significance at p < 0.05 (one-tailed)[8]

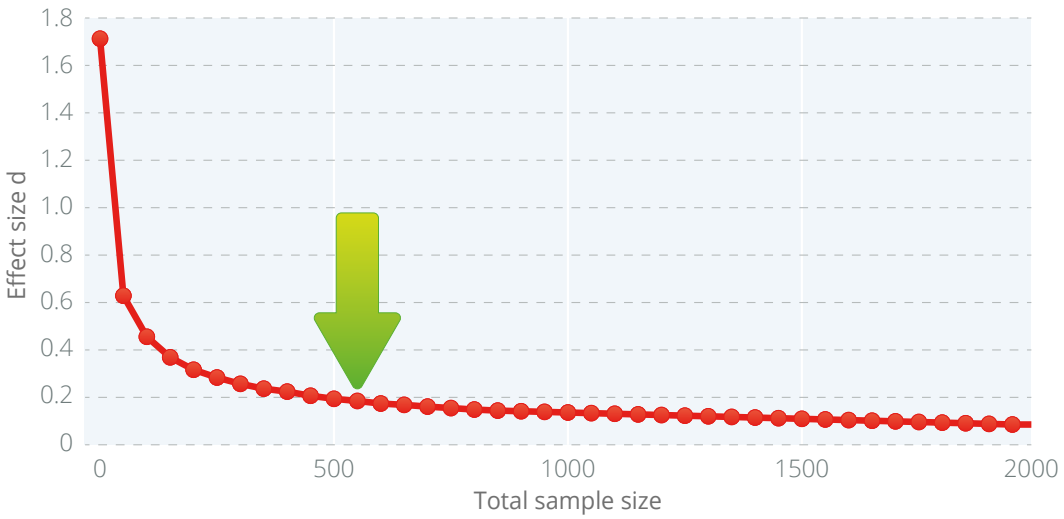
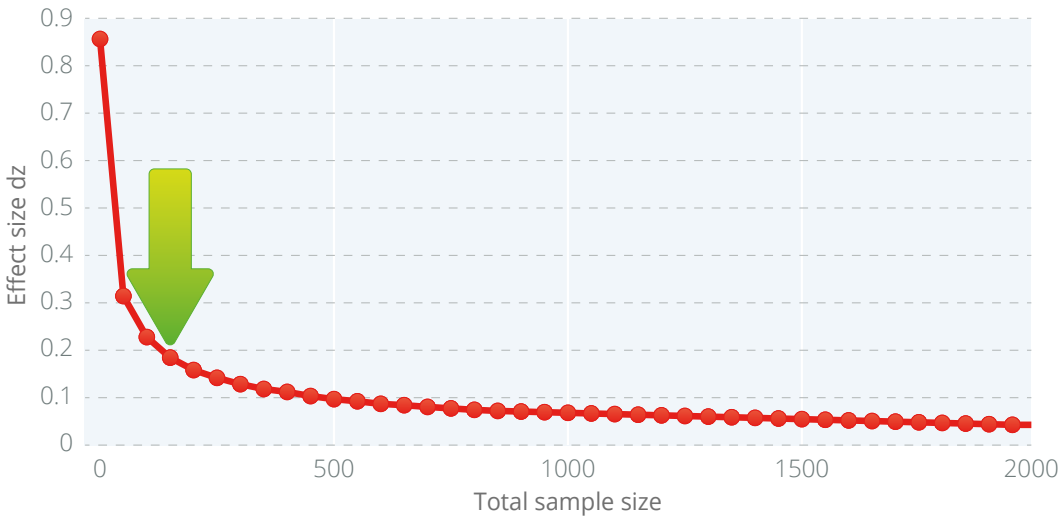


Figure 6: The power to detect an effect between two conditions in a within-subject (or matched pair) design with the threshold for significance at p < 0.05 (one-tailed)[9]



[8] Calculated using G*Power 3.1.7, 80% power to detect an effect with an assumed equal sample size
[9] Calculated using G*Power 3.1.7, 80% power to detect an effect

10

Tonysguboriau Primary School – Ysgol Gynradd Tonysguboriau
Llanhari Primary School – Ysgol Gynradd Llanhari

A small-scale, case-matched, pilot study into the effects of mixed-ability groupings versus ability groupings on pupils' attainment in and enjoyment of numerical reasoning tasks

A small-scale, case-matched, pilot study into the effects of mixed-ability groupings versus ability groupings on pupils' attainment in and enjoyment of numerical reasoning tasks

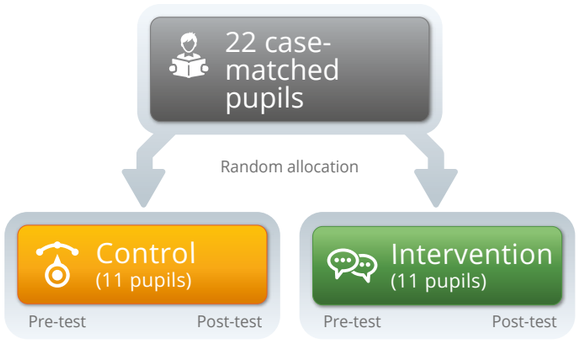
Authors:
Wendy Blyth and Rachel Elphick
wendy.l.blyth.tonysguboriaupri@rctednet.net

Purpose of the research: There is often a debate over groupings of children and how they learn in these groups. Whilst there is always a call for ability groupings and differentiated work with appropriate challenge for procedural learning, we felt that the numerical reasoning skills of some lower-ability children outweighed their procedural ability and therefore could have an impact on the rest of the learners. Through this research, we intended to measure the numerical reasoning progress made by children working in varying groupings (i.e. ability/mixed ability – to see which method of grouping had greater impact on children's learning over a given period of time). The research took place with support from CfBT Education Trust and the National Support Programme in Wales.

The research design

A matched-pair design was used with a pre- and post-test. The independent variable (the use of mixed-ability groups) was defined operationally by creating two conditions:

- IV Level I (Control condition) – Four numerical reasoning activities to be carried out in ability groups
- IV Level II (Experimental condition) – Four numerical reasoning activities to be carried out in a mixed-ability group



Methods

Participants, sample size and randomisation

Two primary schools from the same cluster, with single-form entry, took part in the research. One school undertook the research with a class of 23 Year 6 pupils, whilst the other undertook the research with a class of 30 Year 3 pupils. 22 case-matched pupils from within the classes were randomly allocated to the control or intervention. One pupil did not complete the post-test, therefore analysis was only conducted on the remaining 10 pairs.

Procedures

All children were taught by their own teacher and undertook tasks relevant to their year groups. Tasks were delivered by the class teacher to all the pupils at the same time so that the experimental and control groups received the same input. Activities were delivered in line with normal classroom practice except for the grouping of the children. Teachers from both schools communicated to ensure continuity of approach between the Year 3 and Year 6 classes in the separate schools.

Materials (and apparatus)

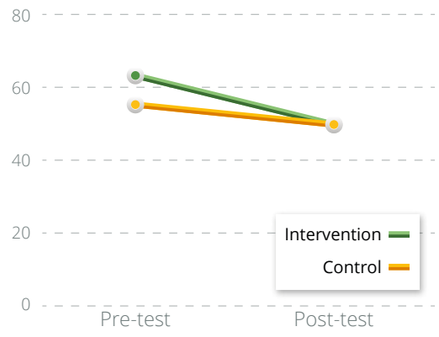
Pre- and post-tests were of a similar nature and were numerical reasoning activities scored from the guidance on the Learning Wales website. A range of numerical reasoning tasks from both the Learning Wales website and NACE were used during the research period. The Enjoyment Gauge (similar to that of a Likert Scale) was also used pre- and post-tests.

Results

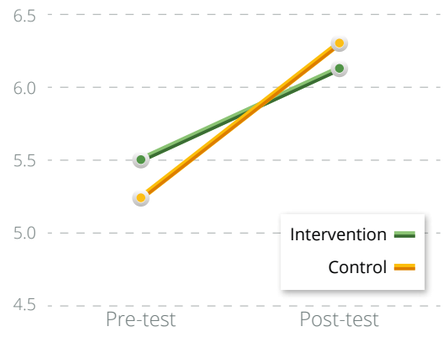
Gain scores were analysed and a Wilcoxon signed-rank test indicated a non-significant ($p = 0.277$ (one-tailed)) effect on attainment for the pupils who were exposed to the mixed-ability grouping ($Mdn = -2.5$) compared to the control ($M = 2.5$). A small negative effect size ($r = -0.125$).

Data from the enjoyment questionnaire was not normally distributed, therefore a Wilcoxon signed-rank test was used. This showed that there was no difference ($p = 0.368$ (one-tailed)) in the levels of enjoyment experienced by the pupils in the mixed-ability group ($Mdn = 1.00$) compared to those in the ability group ($Mdn = 0$).

Pre- and post-test average scores for control and intervention groups



Pre- and post-test enjoyment gauge average scores



Limitations

The small sample size in this pilot study means that the findings must be interpreted with caution. In addition, by necessity the different groups were in the same classes at the same time, therefore cross-contamination may have been an issue, although teachers worked hard to ensure this was not the case.

Conclusions and recommendations for future research

This preliminary study suggests that teachers of Key Stage 2 pupils may be able to have more flexibility in the way that they group children during reasoning tasks than had previously been thought. However, the results should be applied with caution due to the sample size. Enjoyment data suggests that this approach to groupings could be applied without any negative effect on motivation. Replication is recommended due to the small sample size and the possibility that mixed-ability teaching may have a negative effect on attainment. A future larger study may also want to look at the effects of grouping on behaviour.

In partnership with CfBT Education Trust

11

Jumeirah English Speaking School

A six-month mentor programme for underachieving GCSE students in an international school context increases progress across all subjects, as evidenced in GCSE examination results – a non-randomised case-matched study

A six-month mentor programme for underachieving GCSE students in an international school context increases progress across all subjects, as evidenced in GCSE examination results – a non-randomised case-matched study

Author:

Emmet Christopher Glackin
eglackin@jess.sch.ae

Jumeirah English Speaking School

Purpose of the research

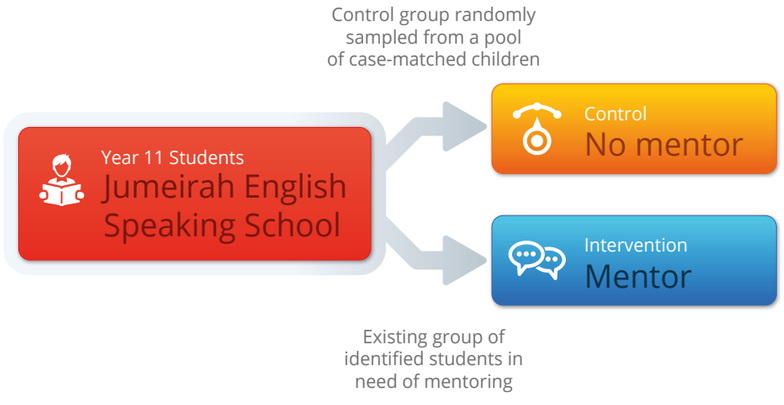
Jumeirah English Speaking School (JESS) is a mixed-ability, not-for-profit, British- and International-curriculum private secondary school in Dubai. The Head of Year 11 has developed an intervention strategy for underachieving GCSE students called the 'Year 11 Mentor Programme'. Students are assigned a mentor whom they meet on a weekly basis to set targets and work on organisational skills. Allocating mentors is resource intensive and therefore it was important to assess the effectiveness of the approach. Prior qualitative research by the present study author suggested benefits – but ones that were mainly affected by ethnicity; therefore it was clear that the study might need to case-match, controlling for this.

The research design

A non-randomised, case-matched design with a pre- and post-test was used. To address the effectiveness of the six-months intervention, with the hypothesis that this would increase actual attainment compared to a baseline of predicted attainment, two conditions were created and defined operationally as follows:

IV Level 1 – Normal teaching practice with no mentor

IV Level 2 – Six months of weekly support with a mentor (experienced teacher)



Progress was measured from students' mock examination results in January against their actual results in August. Case-matching was decided upon based on prior study at Masters level by the researcher.

Methods

Participants, sample size and randomisation

The Year 11 Mentor Programme at JESS consists of 20 students per year receiving support from an experienced teacher. The researcher was able to access historical data from the previous three academic years with a combined sample across three years of 61 Year 11 students. These 61 students were case-matched to a larger pool of similar students who had not been mentored (focusing on ethnicity), from which the control group was then randomly sampled. The total number of students in the final analysis was n = 122. Power analysis (using G*Power) suggested that this sample size would be sufficient (with a threshold for significance of 0.05) to detect a d = 0.23 effect size as significant.

Procedures

Following the school's normal procedures, the Head of Year 11 selected students who had underperformed in their mock examinations in January; this accounted for approximately 15% of the year group. Based upon this information, students in the intervention were assigned a mentor who supported them until their final GCSE examinations. Each mentor was given the same instructions about how to mentor the student. Weekly meetings were held to set targets for progression and correspondence was sent home to update parents of current progress. The Head of Year 11 had an overview of the group and monitored the programme; mentors shared any further concerns where appropriate. Mentors and form tutors knew who was being mentored. However, the students and their subject teachers were kept blind to who was being measured.

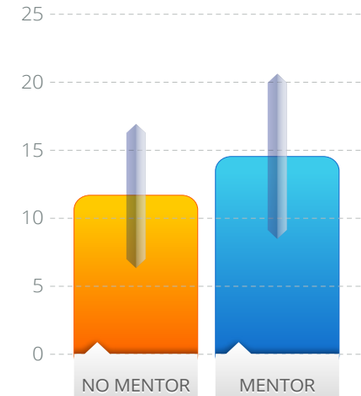
Materials (and apparatus)

Each mentor was given an information pack consisting of instructions that were standardised in terms of content. This pack included weekly report cards that were colour coded depending on progress from targets set in a previous meeting and information from current reporting schedules. Mentors were expected to follow instructions set and keep a record of each student's progress and attitudes to learning.

Results

Gain scores were first calculated from actual and predicted results (examination results minus mock results for each pupil). Statistical significance was set at alpha = 0.05. A paired sample t-test indicated a significant ($p < 0.001$ (one-tailed)) improvement in pupils' progress where they were exposed to the Year 11 Mentor Programme ($M = 14.57$, $SD = 5.94$) compared to the control ($M = 11.67$, $SD = 5.26$) – a moderate positive effect size ($d = 0.514$).

Gain score means for the control and intervention, also showing the standard deviation



Limitations

The sample size for the study was relatively small and the study did not involve random allocation. In addition, the use of case-matching design that focused on one main factor made it hard to fully assess whether between-subject variation may have influenced the findings.

Conclusions and recommendations for future research

The results should be approached with caution because of the sample size and type of design used. However, there is evidence to suggest that the moderately strong improvement in students' progress was not just the result of chance. Recommendations for further research include: increasing the sample size and adding a further measure in relation to the competency of mentors using a Likert scale. An attitudinal survey could also shed further light in relation to students' attitudes to study. Future research may also want to replicate such results with other year groups or focus on particular groups of children (high ability – stretch and challenge) with a fully randomised controlled trial design.

This research was conducted with support from the CfBT Education Trust training programme on Experimental Research Design for Teachers delivered in Dubai in 2015.

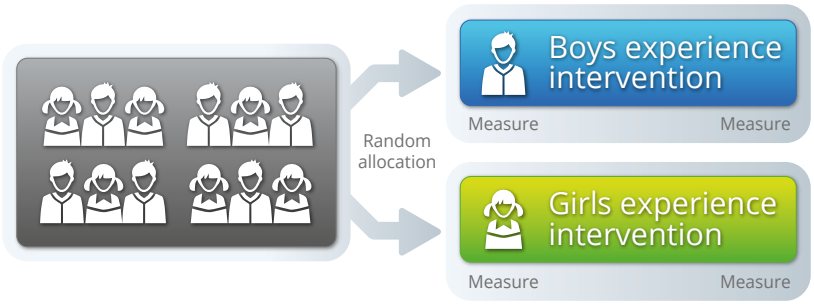
In partnership with CfBT Education Trust

6. Quasi-experimental studies that compare two existing groups

6.1 | Looking at the effects of a single treatment on different groups

It is not always possible to have a control condition, or to randomly allocate. For example, suppose we wanted to compare the effects of a treatment on boys compared to girls. We cannot randomly allocate the participants to being a boy or a girl, because that is what they are. In this case, we can use a form of quasi-experiment in which we accept the existence of the two groups and expose both groups to the same treatment. Figure 7 illustrates the concept.

Figure 7: A quasi-experimental design exploring the effects of a single treatment on boys compared to girls



To remove a degree of bias with regard to who might take place in the study, if we have a large enough population to draw upon, or if we are not testing the whole population group, we might use random sampling to draw down into the study a smaller group of participants, avoiding choosing them directly.

It is also possible to set up your existing groups prior to analysis by identifying who is in what group using some form of measurement. This type of design has been quite common in clinical and psychological research. For example, in hypnosis research it is common to identify which people are low- or high-hypnotic susceptibility individuals (there is a normal distribution in the population) before exposing these two groups to the same form of treatment. Likewise, a researcher might allocate people to parallel groups based on things such as extroversion versus introversion, or high versus low blood pressure.

In the final study in this report, by Jess Moore (a former member of Teach First at the school where the research took place and currently a CfBT Education Trust development analyst) and Simon Andrews, it is this later approach that has been taken. Their report explores whether an attainment gap was closed for low-ability learners compared to average pupils in response to a treatment that all children in the year group were exposed to over a one-year period. Following the use of an initial pre-test of reading age, to identify the lower quartile of reading age pupils and those pupils who would be in the semi-interquartile range (from the 37th to the 62nd percentile around the mean), all children were exposed to the treatment. The researcher then tested the same children at the end of the research period and conducted a range of analyses. They also checked on the upper-quartile pupils to be certain that there were no negative effects for them, bearing in mind the whole year group experienced the intervention.

A final point about this very innovative study is that it demonstrates a possible statistical solution to the challenge of schools exemplifying that they have closed the attainment gap for their low-performing pupils, as well as a means of showing statistically an effective use of additional funding – such as that provided by the Pupil Premium. Following on from the successful application of this approach in the report that follows, an Excel scoring protocol has been devised that specifically deals with the approach. This is available alongside the other Excel spreadsheets that are provided on the CfBT Education Trust training programme Experimental Research Design for Teachers.

What such designs cannot do, of course, is to establish a causal relationship in the way that a study with a control group has the potential to do. Nonetheless, in many school research scenarios it is likely that a school will want to understand the impact of an intervention that has been given to all pupils and the way it makes a relative difference to different groups of pupils; in which case, this type of design is highly suited to such a research question.

12

London Academy

Drop Everything and Read (a one-year reading intervention) closes the attainment gap for a significant number of low-ability Year 7 learners in a zone 5 Academy in London

Drop Everything and Read

(a one-year reading intervention) closes the attainment gap for a significant number of low-ability Year 7 learners in a zone 5 Academy in London

Authors:
Jess Moore
jess.moore@teachfirst.org.uk
Simon Andrews
S.Andrews@londonacademy.org.uk

Purpose of the research: Approximately 60% of students arrive at London Academy with a reading age below their chronological age. Teachers suggested that lower-ability pupils who spend more time reading appear to make more rapid progress. They designed a year-long intervention for Year 7 pupils that brought forward the start of the school day to incorporate a 35-minute Drop Everything and Read (DEAR) session. DEAR was first introduced in September 2013. The school wanted to use a retrospective study to test their hypothesis that the programme closes the attainment gap for a significant number of low-ability Year 7 learners.

The research design

A quasi-experiment design was used with a pre- and post-test to retrospectively assess the effects of a year-long reading intervention (DEAR) on different groups of learners. There were three levels to the independent variable:

- IV Level 1 – Lower attainment learners
- IV Level 2 – Middle attainment learners
- IV Level 3 – Higher attainment learners

Dependent variable – Reading age pre- and post-treatment as measured by the Access Reading Test (Hodder Education)

Methods

Participants, sample size and randomisation

A total of 213 Year 7 pupils were exposed to the treatment. Pre- and post-treatment scores were available for 187 of these pupils and included in the retrospective study. Pupils were assessed using the Access Reading Test to identify the high, middle and lower quartile group of learners. This resulted in 46 lower-quartile learners, 55 middle-band (semi-interquartile range: 37th to 62nd percentile) learners and 47 higher-quartile learners being identified. 39 pupils were identified between the attainment bands and were not included in the study in order to achieve broadly similar sample sizes and a clear definition of pupil category.

Procedures

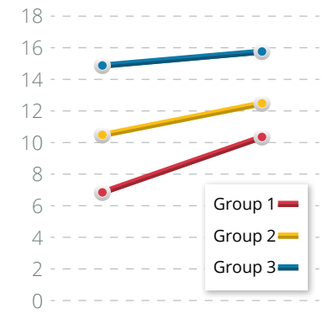
Following completion of the Access Reading Test, all pupils were exposed to the same treatment. The treatment was a Drop Everything and Read (DEAR) reading programme delivered by teachers at the start of each school day for 35 minutes. Pupils were exposed to the treatment consistently across the academic year. Teachers used a myriad of pedagogical techniques to improve pupils' reading ability. Following the programme the pupils were assessed again using the Access Reading Test.

Materials (and apparatus)

Assessment materials used were **Hodder Education's** Access Reading Test forms A and B. Books were selected by practitioners to be appropriate for both age and ability and included novels by Roald Dahl, Michael Morpurgo, Jacqueline Wilson and Suzanne Collins. Practitioners received an initial training session at the start of the programme and met termly to share best practice and learning strategies. The analysis used one of the CfBT Education Trust StatsWizards (Version 10.0).

Results

A 2 x 2 chi-squared test for independence showed a significant number of pupils (42) who had been in the lower quartile on the pre-treatment test were achieving reading age scores equal to the pre-treatment middle band following the intervention ($w = 0.92$ (a very large effect size), $p < 0.001$). Four pupils' attainment remained similar. Thus, 91.30% of previously lower-attaining pupils can be said to have had their attainment gap closed significantly by the intervention. Data for the upper quartile suggests no ill effects for the highest attaining group.



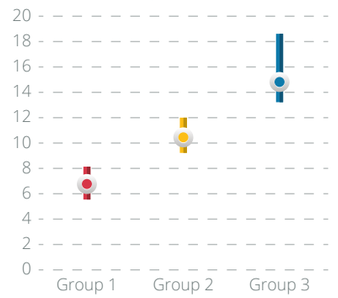
Pre and post-treatment scores as a function of pre-treatment

Group 1: Pupils who were in the lower quartile for the pre-treatment test

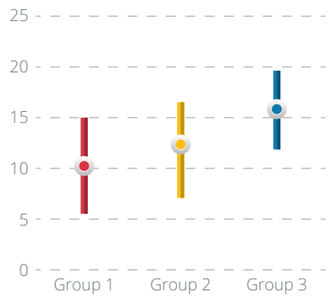
Group 2: Pupils who were in the semi-interquartile range (37th – 62nd percentile) for the pre-treatment test

Group 3: Pupils who were in the upper quartile for the pre-treatment test

Pre-treatment mean and full range of scores



Post-treatment mean and full range of scores



Use of a Kruskal-Wallis test on gain scores also indicated a significant acceleration in progress for the pre-treatment lower band compared to the pre-treatment middle band ($r = 0.35$, $p = 0.002$, a moderately large 'catch-up' effect).

Limitations

It is difficult to isolate the DEAR intervention completely. Other aspects of reading/literacy provision at London Academy may have contributed to the progress in reading age made by lower-ability pupils, for example the increased focus on reading by subject teachers or reading support provided as part of the English curriculum.

Conclusions and recommendations for future research

Results from this retrospective study indicate that the DEAR programme (a one-year reading intervention) closes the attainment gap for a significant number of low- and medium-ability Year 7 learners. Future research may wish to explore whether this same effect can be obtained with a different year group or key stage and/or a longer treatment window, to ascertain if a longer period of intervention delivery results in a greater percentage of learners having their attainment gap closed. An experimental development of this study may better identify the causal factors involved in raising reading attainment using a DEAR-type intervention, and provide recommendations for effective classroom implementation.

In partnership with CfBT Education Trust

CfBT Education Trust
60 Queens Road
Reading
Berkshire
RG1 4BS

0118 902 1000

www.cfbt.com

ISBN: 978-1-909437-69-2 09/2015