A randomised controlled trial of the use of a piece of commercial software for the acquisition of reading skills

Muhammad Khan and Stephen Gorard
School of Education, University of Birmingham, B17 9SX
s.gorard@bham.ac.uk

Abstract

We report here the overall results of a cluster randomised controlled trial of the use of computer-aided instruction with 672 year 7 pupils in 23 secondary school classes in the north of England. A new piece of commercial software, claimed on the basis of publisher testing to be effective in improving reading after just six weeks of use in the classroom, was compared over ten weeks (one term) with standard practice in literacy provision. Pupil literacy was assessed before and after the trial, via another piece of commercial software testing precisely the kinds of skills covered by the pedagogical software. Both the treatment group and the comparison group improved their tested literacy. In a sense the publisher's claim was justified. However, the comparison group improved their literacy scores considerably more than the treatment group, with a standardized improvement of +0.99 compared to +0.56 (overall 'effect' size of -0.37), suggesting that the software approach yields no relative advantage for improvements, and may even disadvantage some pupils.

On this evidence, the use of software, of a kind that is in very common use across schools in England, was a waste of resource. This could be an important corrective finding for an area of schooling that has been the focus of intense policy and practice attention. Of course, the software used has now been superseded by the same and different publishers. But the paper discusses the implications of these results for the use of such software to teach literacy more widely, for the way in which publisher claims are worded, and for the research community in relation to the feasibility of conducting pragmatic trials in school settings.

Keywords

Randomised controlled trial, literacy, reading skills, technology based instruction

A randomised controlled trial of the use of a piece of commercial software for the acquisition of reading skills

Abstract

We report here the overall results of a cluster randomised controlled trial of the use of computer-aided instruction with 672 year 7 pupils in 23 secondary school classes in the north of England. A new piece of commercial software, claimed on the basis of publisher testing to be effective in improving reading after just six weeks of use in the classroom, was compared over ten weeks (one term) with standard practice in literacy provision. Pupil literacy was assessed before and after the trial, via another piece of commercial software testing precisely the kinds of skills covered by the pedagogical software. Both the treatment group and the comparison group improved their tested literacy. In a sense the publisher's claim was justified. However, the comparison group improved their literacy scores considerably more than the treatment group, with a standardised improvement of +0.99 compared to +0.56 (overall 'effect' size of -0.37), suggesting that the software approach yields no relative advantage for improvements, and may even disadvantage some pupils.

On this evidence, the use of software, of a kind that is in very common use across schools in England, was a waste of resource. This could be an important corrective finding for an area of schooling that has been the focus of intense policy and practice attention. Of course, the software used has now been superseded by the same and different publishers. But the paper discusses the implications of these results for the use of such software to teach literacy more widely, for the way in which publisher claims are worded, and for the research community in relation to the feasibility of conducting pragmatic trials in school settings.

Literacy as an issue

Reading is a fundamental skill for later life and forms a basis for any child's subsequent learning at school (Good et al. 1998). Pupils who read well in early stages of their education are more successful in later years compared to those who fall behind (Hirsch 2007). Differential reading ability is a key determinant of patterns of subsequent learning (Wolf and Katzir-Cohn 2001, Pikulski and Chard 2005). Poor reading ability can have harmful psychological, social and economic consequences, with implications far beyond those directly associated with education (Adams and Bruck 1993). Societal demands on reading ability are increasing in the information age (Cunningham et al. 2004), and a minimal level of literacy is an entitlement for all in a civilized nation.

In the UK, concern has been expressed for a decade or more about poor or even, according to some reports, declining levels of literacy. The Department for Education and Employment (1999) famously reported that an estimated 'seven million adults in England cannot locate the page reference for plumbers in Yellow Pages' (p.12). This is clearly an exaggeration. Nevertheless, at least 10% of the children in England used to leave primary school with apparently deficient reading and writing skills (Brooks et al. 2002). Over 30% of children in their first year of secondary school were found not to be able to read at a level suitable for their age in 1997, althouugh this dropped to around 20% by 2009 (National Literacy Trust 2010). At least 30% of 14-year-old boys did not reach the government-specified target level in reading and writing (National Reading Campaign Survey 2005). These apparently high levels of functional illiteracy, for a developed country, are sometimes reflected in the results from international surveys of attainment. For example, England was ranked 19th in the 2006 Progress in International Reading Literacy Study (PIRLS 2007). The figures for low literacy levels differ depending on the age, test used, and the precise standards or targets imposed. It does not even matter for our purposes here whether these claims are true or not. What is clearly true is that policy-makers, educational authorities, many in the media, and some researchers and teachers believe that standards of literacy in schools in England need to improve (ETI 2003, DfES 2003). Can technology-based instruction be part of the solution?

The claims for technology-based instruction

Access to technology in schools in England has grown exponentially since the 1980s. It is now routine for most schools to use technology-based products such as software packages and websites in teaching and learning – both in literacy and other core subject skills. For literacy alone, we estimate that by 2005 there were 300 pieces of software and more than 500 instructional websites available or on the market aimed at improving primary or early secondary reading skills. Part of the reason for this growth has been enhanced government funding for technology-based purchases and for staff development in the use of ICT. Between 1998 and 2002, for example, spending on ICT doubled in secondary schools, and continues to rise. Government bodies in England strongly recommend using technology solutions in all school subjects (DfES 2003).

However, the evidence on the educational benefits of these various technology products is not particularly clear. ICT is used in the classroom, and independently and less formally by individual learners. ICT can be used in the classroom in isolation or as part of a blended learning approach. It may be used for different tasks over different lengths of time, even within one school. The beneficial outcomes sought could be enjoyment, autonomy of learning, future participation, personalization, freeing up teachers to deal with other issues, cost effectiveness, or simply enhanced learning outcomes. It is, therefore, hard to say whether and to what extent technology-based instruction 'works'.

In addition, many of the studies directly addressing the efficacy of ICT in literacy education have been descriptive in nature, relying on the impressions of participants. These studies often find an apparently positive impact on the acquisition of pupil literacy skills (Blok et al. 2002, Silverstein et al. 2000, Cox et al. 2003, Pittard et al. 2003, OFSTED 2004, Baron 2001, Rose and Dalton 2002, Pelgrum 2001, Sivin-Kachala and Bialo 2000). But others have argued that the small sample sizes, the lack of comparators,

indeed the lack of research design, and the passive retrospective nature of some of this work combine to offer a potentially misleading picture (Waxman et al. 2003).

In this light, it is interesting that experimental studies of the effectiveness of software packages in improving literacy skills tend to show rather different results. Rigorous intervention studies with suitable controls often find little or no positive impact from the use of technology-based instruction compared to standard or traditional practice. A number of studies and systematic reviews have found that software packages had no effect on reading achievement (Borman et al. 2009, Rouse and Krueger 2004, Andrews et al. 2005, Torgerson and Zhu 2003, Angrist and Lavy 2002, Goolsbee and Guryan 2005, Dynarski et al. 2007, Lei and Zhao 2005). An overview of reading instruction interventions by Slavin et al. (2008) suggested that mixed methods and co-operative approaches are more effective than technology alone, although this conclusion is itself the subject of some dispute (e.g. Greenleaf and Petrosino 2008).

Given that computers and associated software impose a cost, are frequently updated, and are in widespread use in schools, it is important to have evidence of their impact. The expenditure on ICT may have an impact but not proportionate to the costs. It may have no impact and so be a cost with no benefits. And it may even have deleterious impacts. Untested educational initiatives can frequently be harmful for children (Boruch et al. 2002). 'There are many examples of education interventions that have been widely disseminated on the basis that they are driven by good intentions, seem plausible and are unlikely to do any harm, yet when they have been rigorously evaluated have been found to be ineffective or positively harmful' (Gorard with Taylor 2004, pp.92-93).

This new study examined the claims of a widely used commercial software publisher concerning the impact of a specific literacy program for mainstream year 7 secondary school pupils. At the time of this study, the software was in use in around 400 schools across the UK. Depending on the size of any school, a one year license for this software was between £375 and £600. The publishers claimed that this software was research based, and had been developed with guidance from some of the leading reading experts,

to lead pupils aged 11 to 15 through the essential steps in becoming successful readers. It was considered especially valuable for pupils transferring to secondary school at age 11 or 12, even those with relatively poor reading skills. How did it fare in a pragmatic trial?

Methods used

The participants

Our sample was purposive in the sense that it was regional to Yorkshire in the north of England, and consisted of those state-maintained schools agreeing to co-operate with the research and possessing a minimum level of technology access and support (agreed between the software publishers and the school ICT coordinators). As shown below, the pupils involved in the research were diverse in terms of ethnicity, sex and family poverty, while the schools differed in terms of size, aggregate public examination results and the mix of their pupils. However, because of the replacements necessary we do not claim that these schools and their pupils are statistically representative of a larger known population. The achieved overall sample is the population for this study. This population consists of nine secondary schools in the Bradford, Leeds and York areas. These nine schools agreed that their entire year 7 pupil body would take part in either the treatment or control, as long as the allocation to these groups was for full teaching classes rather than for individuals, and as long as the parents raised no objection. Eight classes out of the total of 31 did not take part because at least one parent objected to their child taking part in the study. This left 23 classes containing 672 pupils at the outset. Four pupils moved to schools in another area before the pre-test, and a further three moved before the post-test. It was not possible to conduct an intention-to-treat analysis using these, since we could not follow the seven missing pupils. Nevertheless, their numbers are small and divided between both groups. A simple sensitivity analysis suggests that their inclusion could make no difference to the clear results of this trial (see below). The resulting 665 pupils ranged in age from 11 years and one month to 12 years and four months.

The 23 classes were ranked in order of size (number of pupils), and a random number generator was used to select one of each successive pair of classes starting with the largest two. The final odd class was randomly allocated to one of the groups. Thus, each class had the same chance of ending up in the treatment or in the control group, once stratified by size. In the end, 11 classes of 319 pupils were in the treatment group selected to use commercial software in their literacy lessons for one term, and 12 classes of 346 pupils were destined to receive standard practice in their literacy lessons over the same period. Table 1 illustrates the outcome of this cluster randomisation. The two groups achieved were very similar in terms of all observed and measured characteristics. The control group had a slightly higher proportion of boys and of pupils not reported as being 'White' in ethnicity. These characteristics may be loosely related to literacy levels in English, and are discussed further below.

Table 1 - Demographic characteristics of treatment and control group pupils in the study

Group	Total	Percentage FSM	Percentage male	Percentage non-White
Treatment	319	21	53	33
Control	346	21	56	34

N=655

Note: FSM represents eligibility for free school meals, an important and widely used measure of family poverty in the UK.

Schools had classes in both groups, and while this might be thought to lead to contamination this was not a strong possibility with this specific intervention (see below), it does lead to some control for a possible school effect. Of course, with only 23 clusters to allocate we cannot be sure that the randomisation had coped with any unobserved systematic differences between the groups. However, the pre- and post-test design (see below) should deal with any revealed differences in literacy levels themselves. Therefore, we should be able to conclude that the major difference between the two groups will be the way in which literacy lessons are conducted for the two groups.

The treatment

The intervention took place for 10 weeks, over a single term, during the course of the academic year 2006/07. The control group remained in routine teaching practice using a more traditional paper and teacher based format, with no specified ICT component.

We do not name here the software used in the treatment (or its publisher). While regrettable, this is what was agreed at the outset, and it anyway makes little difference to the implications of this research (see below). The publishers claimed that their reading software was 'award winning', and that if 11 to 15 year-olds work on this program for one hour a day, spread over six weeks, the program will quickly improve their reading skills including single word reading, sentence reading and non-word reading. It also reportedly improves reading speed, reading fluency, vocabulary, comprehension and reading stamina. The software is multi-sensory in nature, combining vision, sound and touch. It allows pupils to progress at their own pace, with consistent and immediate feedback, and progress is measured. It was designed to be used in conjunction with standard reading exercises, based on the National Literacy Strategy in England. More than 100 starter texts are provided, including poems, tales, recipes, articles, descriptions, letters, points of view, instructions, and official documents, and new content can be added via authoring tools. The standard of difficulty of each exercise, and the look and feel of the program can be adapted to suit the pupil, making it suitable for all ages and abilities including children with learning challenges. It was aligned with National Curriculum standards, developed with guidance from some of the leading reading experts, and grounded in the most current research on literacy, using a carefully structured whole/part/whole approach to reading instruction. It has customised professional development ranging from CD-ROM and online courses to on-site workshops. A comprehensive Teacher's Guide with activities and lesson plans are included in the package. Free technical support is available during office hours.

The number of computers available for use, and access to them, were discussed with the ICT technicians in all participating schools. We found that all the schools had enough

computers and headphones for the treatment group pupils to use during the trial. Trial orientation sessions were made available to all teachers, head teachers, and school governors in the participating schools before the start of the study. A detailed description of the software and its learning activities, the timelines, study purpose, research questions, and expectations relating to participation were provided to all participating teachers. In early September 2006 an agreed final trial procedure was displayed on school notice boards, and a copy of the procedure was provided for all of the teachers involved in the trial. Teachers were provided with log books in which to keep records of implementation, and notes on progress of the trial. To ensure that the treatment implementation was as intended by the software publisher, the researcher tested the software installation on both stand-alone and networked computers in all treatment classes.

Ongoing technical support was agreed with the software publisher for the period of the trial. All treatment teachers received software training about how to use the software from consultants sent by the publisher. The training included a demonstration of the most effective ways of using the software. Teachers were provided with a copy of all materials. They were trained to view individual pupils' feedback, assigned a code for the software publisher's records, and given access to the consultant as well as a telephone number to use if they had any technical problems with the software or with any of the associated activities.

The treatment group used the computer software for a designated time on three to four days each week. In most of the classes, the treatment group was monitored by the teacher while they were working on the computer software. A time was allocated at which pupils were to go to computer laboratories before the intervention started. Headphones were supplied for every pupil to counter distraction, thereby maximising the pupils' attention.

The software, the treatment schedule and the training all encouraged teachers to help pupils complete all of the learning activities provided by the software, over the ten weeks of implementation. The software itself automatically logged the records of each activity completed by each pupil and class. The results appear in Table 2. Most pupils in all

classes completed the bulk of the activities. However, there is some variation, as might be expected in a pragmatic trial. This could be due to differences in ability of the pupils, both in literacy and competence in the use of the software, differences in the length of classes per week, and partly because of the expressed preferences of pupils for some activities over others. The class labeled 1 here had some technical difficulties with their computer system early in the term. This issue is discussed further below.

Table 2 – Percentage of each of 11 classes completing each activity, according to guidelines

Classes	1	2	3	4	5	6	7	8	9	10	11
Activity 1	50	65	96	94	90	100	100	100	100	100	100
Activity 2	45	45	68	96	100	90	79	95	90	90	100
Activity 3	76	76	88	89	97	80	81	79	90	90	100
Activity 4	97	90	90	100	100	81	65	96	90	94	100
Activity 5	79	76	89	80	90	76	100	90	81	100	90

Assessing literacy levels

Both groups were given a pre-test of their existing literacy levels in the first week of September 2006. An equivalent post-test was given to both groups after ten weeks of teaching in December 2006. The assessment was computer-based, perhaps thereby favouring the treatment group slightly, and the items tested were directly linked to the material covered in the treatment software activities. The software used for assessment was the Lucid Assessment System for Schools (LASS secondary, see http://www.lucid-research.com/sales/esales.htm?category_id=31&product_id=183). This looked at eight related reading skills, forming a suite of three attainment tests (single-word reading, sentence reading and spelling), one ability test (reasoning), and four diagnostic tests (auditory memory, visual memory, phonic skills and phonological processing).

Three of the eight tests (sentence reading, spelling, and reasoning) were adaptive, based on statistical item response theory, where each test item is selected from a large bank of

items of known difficulty for 11 to 15 year old pupils. The remaining tests are progressive in format and utilise a graded series of items of increasing difficulty for 11 to 15 year old pupils. For each test, instructions are spoken by the computer, and practice items are given to familiarise the pupil with the test requirements. The LASS software claims to conform to the British Psychological Society's guidelines for the development and use of computer based assessments. Calibration was originally carried out using a representative sample of UK pupils, with ages ranging from 11 to 15 years, and subsequently with pupils aged 8 to 11. There is a reasonable correlation between the LASS tests and the NFER Sentence Completion Test of reading comprehension. Each of the tests shows reasonably high internal consistency between items (around 0.9), and also test-retest reliability (around 0.75).

The LASS software generated test results in raw and z-score form, with percentiles and age-expected equivalents. The z-scores were converted to standard scores, by adding 100 to 15 times the z-score. The total of the eight standard scores, with an expected mean of 800, were then used to calculate changes over time or between groups.

Further data collection

For any research design, but perhaps especially for a trial, it is important to collect more than the presumed outcome data (Gorard with Taylor 2004). In-depth and contextual data can help explain why an intervention does or does not work, how to improve it, or which sub-groups of learners it is most appropriate for.

Before the trial, we gathered the age, sex, free school meal eligibility and ethnic origin of all pupils involved (as these measures would appear on the pupil-level annual schools census in England).

A template was developed for the study in conjunction with the software publisher, in order to record classroom observations. This template was used by teachers to summarise pupil learning activities, their use of software, and the general environment of the ICT

rooms. Implementation logs were developed for the treatment teachers to record the pupil activities (such as games, activities, units and sections covered in their ICT laboratories and classes), and information about the way in which activities were delivered and any technical problems associated with the software. The teachers were requested to log the information regarding the number of software activities started and completed by the pupils, the type of software learning activities used by the pupils, the role of the teachers during the treatment session, and the number of pupils absent from the sessions. The research team visited all schools on a regular basis during the 10 weeks, and ensured that no other instructional software was used in the treatment group classes.

Unstructured discussions about the trial, lasting 20 to 30 minutes, took place with teachers, head teachers, and school governors. Their chief purpose with the treatment group was to find out if the software was considered effective in improving reading skills and, if so, why it was effective and what actually made it effective. All respondents also discussed the attitude of pupils towards using the software. For the control group one of the purposes was to find out about any other software used.

Analysis

In this paper, we present the overall results from the pre- and post-tests of reading skills. These are the mean (and standard deviation) scores for each group. We present a standardised improvement for each group, calculated as the gain from pre- to post-test divided by the overall standard deviation at the pre-test. Some concerns have been expressed at the use of this standard deviation in cluster trials (White and Thomas 2005). However, this concern is about comparability with individual trials in a meta-analysis – a rather esoteric topic – and anyway only serves to warn us that the improvement of both groups might be underestimated. As is made clear below, such an underestimation for both groups would make no difference to the substantive findings here. We also calculate the standard 'effect' size as the difference between the gain scores for the treatment and control, divided by the pooled (average) post-test standard deviation of both groups. Although our calculation does not, in itself, identify any cause:effect model,

our clear intention was to design the study so that any major difference between the groups could be attributed to the intervention.

We make mention of other results, such as the gains and effect sizes for sub-groups of pupils such as boys and girls separately. We also present an overall picture of the unstructured interview and observation data. These will all be dealt with more fully in another paper. For more on trial design and analysis see the resources available via http://www.tlrp.org/capacity/rcbn.html.

The overall findings

At the outset of the trial, the pre-test scores show that both groups had similar standardised mean scores and deviations (Table 3). Both means were above the expected mean score of 800 (see above). The treatment group was very slightly superior at this stage. After 10 weeks of software use in literacy lessons the treatment group improved their standardised mean score substantially, just as the software publishers had claimed. Therefore, a simple before and after design with no control could easily, but falsely, conclude that the use of commercial software was an especially effective approach to literacy teaching and learning. This illustrates again the danger for educational research of conclusions drawn from what constitutes the majority of published work, conducted without suitable comparators (see examples in Gorard et al. 2007).

Table 3 – Pre- and post-test scores for both groups

	Pre-test mean	Pre-test SD	Post-test mean	Post-test SD
Treatment	823	68	863	88
Control	817	72	886	78

N=655

As Table 3 illustrates, the control group exposed to standard practice in literacy lessons, no routine access to computers, and no access to the treatment software, also improved

their standardised mean score substantially. There is no *prime facie* case here that the improvement for the treatment group was due to the software used. In fact, it would be easier to mount an argument that pupils using the software were disadvantaged. Insofar as the methods used in this study are accepted, we have shown that the software package used was ineffective in comparison to standard practice. This is an important conclusion, with wider implications than might be imagined at first sight. The implications are discussed at the end of the paper. Another noticeable finding in Table 3 is the increase in the standard deviation relative to the mean score for the treatment group at post-test compared both to pre-test and to the control group. This means that the eventual results for the treatment group were more varied, and raises the possibility that the treatment was less effective for only some pupils.

The gain score for each group in the trial is converted to a standardised improvement, as described above. This shows even more clearly the greater improvement for pupils in the control group (Table 4). The precise figure used as the standard deviation for calculating the effect size does not matter. The substantive findings are the same whether the overall standard deviation (71) is used, as here, or whether each group uses its own pre- or post-test standard deviation (see Table 3). In all cases, the improvement score for the control group is nearly twice as large as for the treatment group. This translates into an effect size of -0.37 (or 30/81).

Table 4 – Gain scores for both groups

	Gain score	Overall pre-test SD	Standardised improvement
Treatment (319)	40	71	0.56
Control (346)	70	71	0.99

N=655

Intriguingly, the in-depth data collected routinely as part of the trial suggested a high level of satisfaction with the treatment. The technology-based instruction reportedly provided teaching groups with a range of information, links and activities in an accessible

and entertaining way. Nine of the 11 teachers involved in the treatment said that the software had an encouraging focus on language for early Key Stage 3 and that the activities were stimulating for pupils and teachers alike. They believed that it offered a reliable way to help pupils improve their reading skills. The pupils were satisfied with the technology-based reading materials, and were observed getting heavily involved in the activities. When asked, all teachers indicated that they would use the same or similar software in the future, and almost all of them said that they would recommend it to other teachers.

One limitation that was repeated by teachers was that the software did not cover everything the teachers would have wanted to cover in the term it was used. Several teachers reported minor technical problems with the software or more often with their school computer systems, and one class (class 1 above) apparently wasted several lessons because of a school computer network failure.

Discussion

The trial design eliminates several alternative interpretations of our results but, as with any piece of research, it is important to consider the warrant for our claim that the software was ineffective in comparison to standard practice (Gorard 2002a). If that conclusion were not, in fact, true how else could we explain our findings?

The class that had extended difficulty with their computer system had a low gain score which contributed in part to the poor showing of the treatment to the control. However, it does not explain all of the difference. It certainly does not suggest that the treatment was more effective than the control. And in a pragmatic trial we expect these differences. Such technical difficulties are a real consideration for technology-based instruction and should not be 'cleaned' away to enhance the apparent effect of the treatment, any more than poor teaching of the control would be ignored on the basis that it 'should not happen'.

The sample of 23 classes with 655 pupils could be considered quite small, although the trial is one of the larger ones of its kind in the UK. Brooks et al. (2006), for example, had an achieved sample of only 130 pupils. Perhaps it was not of the scale (did not have the 'power') to demonstrate the advantage of the software. But there are two reasons why we suggest this is not so. First, as our results show this is not really an issue of power. It is not that the treatment group did slightly better but that we cannot be sure whether the difference is statistically 'significant' (with all of the flaws inherent in such a claim, Gorard 2010). The treatment group did substantially *worse*. Second, the sample size is larger than most studies of this type, and is an order of magnitude higher than the samples used by the publisher to test the software initially. Third, and most obviously, there is an asymmetry in the burden of scientific proof. The publishers claim that the software will be more effective than standard approaches (else why would anyone buy it?). We tested it and found no evidence that it was more effective. Until someone does another controlled study with randomisation of treatment but showing a different result, we must assume that this treatment is ineffective.

Schools had classes in both the treatment and control conditions, and this can lead to 'contamination' whereby treatment pupils share their new approach with control pupils, so reducing the apparent effect size of the treatment. This is a standard concern in cluster randomised trials using teaching groups, and is partly handled by intention to treat analysis. More specifically, the treatment was based on content presented using the technology, was not made available in the system to other pupils (login name and password-protected), and it is unlikely that the treatment pupils would remember the details of the activities or discuss them in detail during free time. It is not like passing over a pill, or lending a textbook. Contamination is still possible, but this could not explain the substantially *lower* level of progress made by the treatment group. The intervention was also more natural than in some prior studies, such as Brooks et al. (2006) where hardware shortages meant some distortion of the way in which pupils were grouped in classes.

The control group started from a slightly lower base score. In some tests this would give them more room for improvement. This could perhaps explain some of the greater gain, but does not explain the larger absolute score for the control in post-test. Pupils declaring a non-White ethnicity were very slightly more prevalent in the control, and showed higher gains in both groups. But the scale of these small differences cannot explain the results. If non-White pupils are excluded from analysis the result stands. There were more boys than girls in the control, but their gains were equivalent. Only boys in the treatment group showed higher gains than girls. This is an interesting result, perhaps due to boys' expressed preference for using ICT in lessons, and we will discuss the in-depth results in more detail in future papers. For the present, we simply note that the group-level result works against explaining why the treatment group showed such small gains.

Our result depends heavily on the validity of the test. If this did not test the learning outcomes fairly, then our conclusion is in doubt. We conducted lengthy face validity checks, comparing the material in the tests with that in the exercises and found a good match. We were interested to note that our result would hold for any one of the eight individual tests (such as spelling) as well as for the overall scores reported here. And the fact that the test was computer-based should not disadvantage the treatment group more used to handling such material on the computer (rather the reverse). Finally, and perhaps more importantly for a pragmatic trial, the software publisher accepted the test as valid before the start.

The test, and the trial it supports, is only intended to take into account the eight reading skills involved. There may be unmeasured gains in other respects – beneficial unintended consequences perhaps. But there might also be unmeasured harm caused by the treatment. We do not know. We could, of course, have conducted other trials – perhaps comparing standard classes to a blended approach of ICT and traditional teaching, or comparing the software classes to having no classes at all. All of these might generate different results. But what are the implications of what we did find?

Implications

The simplest and most obvious implication of the study is that this software, for these classes, at this stage, was not effective. In fact, it was markedly less effective than standard classroom practice not involving technology. Unfortunately, like most reasonably secure claims in the social science of policy or practice, this implication is of limited practical value since it is retrospective. We are naturally in sympathy with the demand of Brooks et al. (2006) that 'Before schools adopt ICT packages for teaching spelling or other literacy skills, these need to be evaluated in randomised controlled trials' (p.142). But we do not think this is always feasible. Our fieldwork was conducted in the academic year 2006/07, and at time of writing the specific version of the software used in the trial is no longer marketed. With ICT and education, in particular, Moore's law means that once a longitudinal study is complete the technology tested will have been superseded. By the time this paper is published, even those schools which had already purchased the precise version of the software tested will likely have moved to a new product. There is no obvious way around this limitation that would also apply, for example, to rigorous evaluation of a national policy. Does this mean, then, that we should not bother to conduct longitudinal evaluations of this kind?

We would argue that evaluations such as these have a wider purpose, and more general implications. Most interventions in education are not rigorously evaluated before implementation. Where they are evaluated *post hoc*, many are found to have been ineffective or worse (see Gorard 2006 for examples). It is then too late for any such intervention, and the missed opportunities and even harm they may have caused to learners. But as in the fuller design science process (Gorard and Cook 2007), we can use what we have learnt about the nature of effective and ineffective interventions to help design better ones for the future, to make formative 'corrections' to live implementations, and to provide clear guidelines about the kinds of claims that can legitimately be made about any educational artefact (like a piece of software). We suggest some of these in the following paragraphs.

Many educational programs have in-built assessment and record-keeping applications. Software publishers then make their marketing research claims based on these records. The problem is that in most of the cases these findings are not based on a comparison with anything else. So, for example, the software publisher claims about the effectiveness of their product should be tempered by a caution that no suitable comparator was used, or perhaps all such claims should only be allowed by advertising standards when a suitable comparator has been used. Otherwise, teachers and educational authorities are in danger of being duped by a claim that is the same in real terms as the kind that is rightly banned in the medical literature. For example, since the common cold is of short duration we are not allowed to advertise a pill that is 100% effective in clearing up colds after one week.

Another important implication is the warning against over-reliance on impressions about the treatment efficacy, as reported by people involved in any intervention. The overwhelming view of staff and pupils involved in the treatment group was that learning was proceeding well, that pupils were better motivated and enjoying lessons, and teachers had been freed to deal with specific literacy difficulties within their classes. These reports are valuable in themselves, and open up possible areas to investigate ways in which technology-based instruction could lead to other beneficial outcomes. But the reports do not agree with the results of the central testable claim made by the publishers that use of the software will lead to enhanced performance in assessment of literacy skills after just six weeks. If our results are accepted, then many of the impressions of staff and pupils about the efficacy of the product were incorrect, just as the publisher claims were found to be incorrect.

More generally, the results of this study add to the body of research that shows concerns about the effectiveness of technology-based instruction. Several prior studies have shown limited or no beneficial impact of ICT-based approaches on pupil learning in reading skills (as here) and in other core subjects such as mathematics (see above). But in isolation these studies do not reveal the scale of wasted opportunities and possible harm done to pupils. In general, marketing teams in the UK offer software to schools on a trial basis. During the trial they show how pupils are making progress by using the in-built

assessment process (without appropriate comparator). Teachers can then see pupil progress over learning activities and may be persuaded to purchase. Once teachers have bought the software they tend to use the convenient in-built assessment process regularly. This makes all involved part of a reinforcing cycle. The software publishers make money. Teachers have a record of progress made by pupils, for their own and others satisfaction. Pupils generally enjoy working on computers and playing with different technology applications. Parents will be pleased to see a record of their child's progress. Local and national government is content that their funding of technology initiatives is justified, and schools are persuaded to spend that funding on technology products, making money for the companies to develop new products.

Our final suggested implication concerns the future of evaluation research. This new study demonstrates again the feasibility of classroom-level randomized controlled trials (and similar active rigorous research designs). It is possible for a small team or even a lone researcher to conduct a trial of this scale within one year. The cost of this unfunded study was minimal, mostly related to travel between schools for the fieldwork. Experimental designs are not inherently expensive. Most of the cost of an intervention study is for the intervention, which means that pragmatic evaluation of an intervention that is already scheduled to take place costs almost nothing additionally. No special ethical or practical issues arose. There was a *prime facie* case for implementing the intervention, which would have taken place in schools without the research anyway. It cannot be unethical to deny the intervention to the control when the treatment was inferior anyway. And it cannot be ethical to permit an inferior and costly treatment to be rolled out without testing (Gorard 2002b). Relatively small trials such as this one are an important part of the mixed methods necessary to conduct education research. If education researchers really care for their ultimate charges (the learners) then, as part of their wider approach, they will embrace this simple, cheap and ethical way of evaluating the many interventions taking place routinely in schools.

Acknowledgements

Thanks go to all schools agreeing to participate in this study. Special thanks go to the publishers of the literacy software and assessment software used, for their generosity in supporting this study through provision of software and with technical support throughout.

References

Adams, M. and Bruck, M. (1993) Word recognition: The interface of educational Policies and scientific research, *Reading and Writing: An interdisciplinary Journal*, 5, pp. 113-139

Andrews, R., Dan, H., Freeman, A., McGuinn, N., Robinson, A. and Zhu, D. (2005) *The effectiveness of different ICTs in the teaching and learning of English (written composition) 5–16*, Research Evidence in Education Library, London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London, http://eppi.ioe.ac.uk/EPPIWeb/home.aspx?&page=/reel/reviews.htm, accessed 26/3/06

Angrist J. and Lavy, V. (2002) New evidence on classroom computers and pupil learning, *The Economic Journal*, 112, pp. 735–765

Baron, D. (2001) From pencils to pixels: the stages of literacy technologies, pp. 70-84 in *Literacy: A Critical Sourcebook.* Boston, MA: Bedford/St. Martin's

Blok, H., Oostdam, R., Otter, M., and Overmaat, M. (2002) Computer-assisted instruction in support of beginning reading instruction: a review, *Review of Educational Research*, 72, 1, 101-130

Borman, G., Benson, J. and Overman, L. (2009) A randomised field trial of the Fast ForWord Language computer-based training program, *Educational Evaluation and Policy Analysis*, 31, 82-106

Boruch, R., De Moya, R. and Snyder, B. (2002) The importance of randomised field trials in education and related areas, pp. 50-79 in F. Mosteller and R. Boruch (Eds.)

Evidence Matters: Randomised Trials in Education Research, Brookings, Washington DC

Brooks, G., Cole, P., Davies, P., Davis, B., Frater, G., Harman, J. and Hutchison, D. (2002) *Keeping Up with the Children*, Evaluation for the Basic Skills Agency by the University of Sheffield and the National Foundation for Educational Research. London: Basic Skills Agency

Brooks, G., Miles, J., Torgerson, C. and Torgerson, D. (2006) Is an intervention using computer software effective in literacy learning? A randomised controlled trial, *Educational Studies*, 32, 2, 133-43

Cox, M., Abbott, C., Webb, M., Blakeley, B., Beauchamp, T. and Rhodes, V. (2003) *ICT and pedagogy, A review of the research literature*, ICT in Schools Research and Evaluation Series No. 18, Coventry/London: Becta/DfES, http://www.becta.org.uk/page_documents/research/ict_pedagogy_summary.pdf, accessed 10/2/05

Cunningham, M., Kerr, K., McEune, R., Smith, P. and Harris, S. (2004) *Laptops for Teachers, an Evaluation of the First Year of the Initiative,*, ICT in Schools Research and Evaluation Series No. 19. Coventry/London: Becta/DfES, http://www.becta.org.uk/page_documents/research/lft_evaluation.pdf, accessed 18/12/06

Department for Education and Employment (1999) *Radical change needed to boost basic skills. A briefing paper on the report 'A Fresh Start - Improving Literacy and Numeracy'*, Skills and Enterprise briefing, Issue 5/99, London: Department for Education and Employment

DfES (2003) *Towards a unified learning e-learning strategy*, London: HMSO, http://www.dfes.gov.uk/consultations/downloadableDocs/towards%20a%20unified%20e-learning%20strategy.pdf, accessed 22/3/05

Dynarski, M., Agodini, R., Heaviside, S., Novak, T., Carey, N., Campuzano, L., et al. (2007) *Effectiveness of reading and mathematics software products: findings from the first pupil cohort*, (Publication No. 2007-4005), Washington, DC: U.S, Department of Education, Institute of Education Sciences, available from http://ies.ed.gov/ncee/pdf/20074005.pdf

ETI (2003) *An evaluation by the education and training inspectorate of information and communication technology in post-primary schools 2001–2002*, Education and Training Inspectorate, http://www2.deni.gov.uk/inspection_services/surveys/index.htm, accessed 9/1/05

Good, R., Simmons, D., and Smith, S. (1998), "Effective academic interventions in the United States: Evaluating and enhancing the acquisition of early reading skills." *School Psychology Review*, vol. 27, pp. 45-56.

Goolsbee, A. and Guryan, J. (2005) The impact of internet subsidies for public schools, *Review of Economics and Statistics*, 88, 2, 36-347

Gorard, S. (2002a) Fostering scepticism: the importance of warranting claims, *Evaluation and Research in Education*, 16, 3, 136-149

Gorard, S. (2002b) Ethics and equity: pursuing the perspective of non-participants, *Social Research Update*, 39, 1-4

Gorard, S. (2006) Does policy matter in education?, *International Journal of Research and Method in Education*, 29, 1, 5-21

Gorard, S. (2010) All evidence is equal: the flaw in statistical reasoning, *Oxford Review of Education*, iFirst, pp.1-15

Gorard, S. and Cook, T. (2007) Where does good evidence come from?, *International Journal of Research and Method in Education*, 30, 3, 307-323

Gorard, S., with Adnett, N., May, H., Slack, K., Smith, E. and Thomas, L. (2007) *Overcoming barriers to HE*, Stoke-on-Trent: Trentham Books

Gorard, S., with Taylor, C. (2004) *Combining methods in educational and social research*, London: Open University Press

Greenleaf, C. and Petrosino, A. (2008) Response to Slavin, Cheung, Groff and Lake, *Reading Research Quarterly*, 43, 4, 349-354

Hirsch, D. (2007) *Chicken and Egg: child poverty and educational inequalities*, London: Campaign to End Child Poverty, http://www.endchildpoverty.org.uk/index.html, accessed 14/9/07

Lei, J., and Zhao, Y. (2005) Technology uses and pupil achievement: A longitudinal study, *Computers and Education*, 49, pp. 284–296

National Literacy Trust (2010) Are children's literacy skills improving or getting worse, http://www.literacytrust.org.uk/about/faqs/filter/about%20literacy%20in%20the%20uk#q713, accessed July 2010

National Reading Campaign Survey (2005) *Making every home a reading home*, http://www.literacytrust.org.uk/press/FRC.html, accessed 28/2/06

OFSTED (2004) *ICT in schools: the impact of government initiatives*, School Portraits Eggbuckland Community College, London: Ofsted, www.ofsted.gov.uk/publications/index.cfm?fuseaction=pubs.displayfile&id=3704&type=pdf, accessed 26/3/06

Pelgrum, W. (2001) Obstacles to the integration of ICT in education: results from a worldwide educational assessment, *Computers and Education*, 37, pp. 163-178

Pikulski, J. and Chard, D. (2005) Fluency: Bridge between decoding and comprehension, *The Reading Teacher*, 58, 6, 510-519

PIRLS (2007) PIRLS 2006 reading achievement, http://news.bbc.co.uk/1/hi/education/7117231.stm, accessed 3/9/09

Pittard, V, Bannister, P and Dunn, J (2003) *The big pICTure: The impact of ICT on attainment, motivation and learning*, London: DfES, http://www.dfes.gov.uk/research/data/uploadfiles/ThebigpICTure.pdf, accessed 22/11/05

Rose, D. and Dalton, B. (2002) Using technology to individualize reading instruction, pp. 257-274 in C. Block, L. Gambrell and M. Pressley (Eds.) *Improving comprehension instruction: Rethinking research, theory, and classroom practice*, San Francisco: Jossey Bass Publishers

Rouse, C., and Krueger, A. (2004) Putting computerized instruction to the test: A randomised evaluation of a "scientifically-based" reading program, *Economics of Education Review*, 23, pp. 323–338

Silverstein, G., Frechtling, J. and Miyoaka, A. (2000) *Evaluation of the use of technology in Illinois public schools: Final report* (prepared for Research Division, Illinois State Board of Education), Rockville, MD: Westat

Sivin-Kachala, J., and Bialo, E. (2000) *2000 research report on the effectiveness of technology in schools* (7th ed.), Washington DC: Software and Information Industry

Slavin, R., Cheung, A., Groff, C. and Lake, C. (2008) Effective reading programs for Middle and High Schools: A best-evidence synthesis, *Reading Research Quarterly*, 43, 3, 290-322

Torgerson C. and Zhu D. (2003) *A systematic review and meta-analysis of the effectiveness of ICT on literacy learning in English, 5-16,* Research Evidence in Education Library. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London

Waxman, H., Lin, M., and Michko, G. (2003) *A meta-analysis of the effectiveness of teaching and learning with technology on pupil outcomes*, North Central Regional Educational Laboratory Web site: http://www.ncrel.org/tech/effects2/waxman.pdf, accessed 28/2/06

White, I. and Thomas, J. (2005) Standardized mean differences in individually-randomized and cluster-randomized trials, with applications to meta-analysis, *Clinical Trials*, 2, 141,151

Wolf, M. and Katzir-Cohen, T. (2001) Reading fluency and its intervention, *Scientific Studies of Reading*, 5, 3, 211-239