Journal of Research in Reading, ISSN 0141-0423
Volume 26, Issue 3, 2003, pp 267–279

Computerised formative assessment of reading comprehension: field trials in the UK

Keith J. Topping and Anna M. Fisher

Faculty of Education and Social Work, University of Dundee, Scotland

Increased curriculum time allocated to reading might not be effective in raising achievement. Teachers need to closely monitor and manage both the quality and quantity of individualised reading of all their pupils for optimal effectiveness. 'Learning information systems' (LIS) for reading such as the 'Accelerated Reader' (AR) enable this through individualised computerised assessment of pupil comprehension of 'real books', with feedback to both pupil and teacher. This study explored the impact of AR on reading achievement in 13 schools of different types spread across the UK, the majority socio-economically disadvantaged. Participating pupils were aged 7–14 years. Pre-post norm-referenced gains in reading achievement were measured by group paper-reading tests and a computer-based adaptive reading test. The implementation integrity of AR was assessed by direct observation by researchers and through data generated by the programme itself. On both paper and computer-based reading tests, on aggregate pupils in the 13 schools gained in reading at abnormally high and statistically significant rates. Boys tended to show larger gains than girls on the paper test. However, implementation integrity was very variable. In particular, some teachers failed to intervene in response to AR data indicating that pupils were reading ineffectively. AR appears to have potential for raising reading achievement, but only if implemented appropriately.

Monitoring reading practice

Many studies have found high positive correlation between levels of reading practice (at school or home) and reading achievement (e.g. Allington, 1984; Anderson, Wilson & Fielding, 1988; Biemiller, 1978; Donahue, Voelkl, Campbell & Mazzeo, 1999; OECD, 2002; Rowe, 1991; Snow, Burns & Griffin, 1998; Stanovich, 1986; Topping & Paul, 1999). A smaller number of studies (e.g. Leinhardt, 1985; Shany & Biemiller, 1995; Taylor, Frye & Maruyama, 1990) evidenced a causal direction from practice to achievement, rather than vice versa or confounded.

However, simply increasing time allocated to reading practice might not be effective in raising achievement. Indeed, a review of the practice of allocating sustained silent

reading (SSR) time in schools noted mixed results, six studies finding a positive effect on reading scores, but five no such effect (Manning-Dowd, 1985). 'Reading practice' is not a homogeneous, unitary activity, and the quality and effectiveness of reading practice also requires consideration. Arguably, pupils need to practise reading at a level at which they are appropriately challenged by exposure to new vocabulary and concepts, but not confronted with failure, avoiding unproductive reading at levels too low or high for effective learning to take place.

The practical problem for the class teacher is how to monitor the day-to-day individualised reading of all their pupils, to check that such activity is optimally effective and to facilitate pedagogical intervention to shape it towards effectiveness. Computerised 'learning information systems' (LIS) for reading seek to provide teachers with a tool to enable them to achieve this otherwise daunting task. One such tool is known as the 'Accelerated Reader'. (Note that learning information systems are not to be confused with 'integrated learning systems', which deliver both curriculum content and assessment related to that content.)

What is the Accelerated Reader (AR)?

The Accelerated Reader (AR) (Advantage Learning Systems, 1993) is a system for free-standing computer assisted individualised assessment of comprehension of 'real books'. Pupils select individualised books from the many thousands of titles for which AR quizzes are available, and read at their own pace, at school and at home. On completion, they take the multiple-choice AR comprehension test for the book at the computer. Each book has a maximum point value according to its length and difficulty. When the pupil self-tests, the programme awards points up to this maximum, according to their number of correct test responses.

Teachers may choose to allow pupils to test on books read to and with them, as well as those read independently and silently, especially in the case of early or delayed readers. Where the programme is used with class-wide, selective or elective peer tutoring, both assisting and assisted participants may subsequently self-assess their comprehension of the book (Topping, 2001). For emergently literate tutees, the most recent version (AR Universal) provides quizzes with digitised speech output. The default condition is that pupils can test on a book only once.

The AR quizzes are brief, primarily assess literal comprehension rather than idiosyncratic reader inferences or other more complex responses (which might be culturally specific), and do not measure all relevant reading behaviours in school or elsewhere. An AR quiz typically has 20 items, each item posing a question and offering four response options. (For example, one question and response options for the book *Born Free*: why do elephants sometimes kill lions? A. They regard lions as the only enemies of their young; B. Elephants often kill animals that encroach on their territory; C. Lions deprive elephants of their food supply; D. When an elephant is cornered, it becomes very aggressive.) Indeed, AR questions are deliberately restricted to those that demonstrate adequate psychometric reliability. Consequently, AR points gained are likely to be a consistent and accurate measure of the quantity and difficulty of words read and comprehended, and therefore a useful aggregate measure of successful reading practice.

For the assessment of idiosyncratic reader inference in a more open-ended way, teachers might generate their own questions, have pupils who have read the book generate questions for each other, or use the 'literacy skills' tests extension of AR. This

latter is available for a smaller number of titles, assessing and reporting on 24 generic higher-order literacy skills, including inferential reasoning, main idea, cause and effect, characterisation and recognising plot (but does not claim the reliability and validity of the regular reading practice quizzes).

The AR programme provides the teacher with an automatically updated analysis of scores for individuals, whole classes, or other groupings. This indicates average percentage correct, difficulty of books read, points earned and other diagnostic information. The software designers recommend that teachers target a quiz success rate of 85% correct as optimal for pupils, with either independent or supported reading. Automatically computer-generated 'at risk' reports flag a need for the teacher to intervene with any pupil whose reading activities appear currently ineffective. This might include pupils reading at high as well as low levels. Further detail of how AR works is available on-line in Topping (1999).

The software originated in the USA, where AR is currently in almost half of the schools. This paper reports the first extensive field trials of the software in the UK.

Theoretical underpinnings

The AR programme is intended to impact on learning effectiveness by providing structured and detailed feedback: directly to the pupil; directly to the teacher; and to the pupil mediated and interpreted by the teacher.

The expectation is that this feedback will lead to adjustment and promote more effective subsequent performance (Bangert-Drowns, Kulik, Kulik & Morgan, 1991; Butler & Winne, 1995; Lhyle & Kulhavy, 1987). Kluger and DeNisi (1996) meta-analysed the effects of feedback interventions on performance, concluding that feedback could yield a significant positive effect on performance, and that computer feedback was associated with particularly large effect sizes. Goal clarity and pupil commitment and belief in success were important moderating influences on effectiveness.

Considering impact at a very simplistic level, the quiz helps ensure the pupil has actually read the book, since reading choices are individualised and any attempt to copy quiz responses from another reader both difficult and pointless. More subtly, for the pupil the points give timely feedback on the successfulness of reading, in terms of basic comprehension. This should enable greater pupil control over reading activity, in terms of management of appropriate challenge and other parameters of the reading process (Shapiro & Cole, 1994). Through this, meta-cognitive awareness might be heightened and feelings of self-efficacy as a learner enhanced (Schunk, 1994).

For the teacher, detailed feedback on the reading performance of all pupils in the class is provided without expenditure of teacher time. An indication of the successfulness of each pupil's reading performance is available, in relation to each pupil's current reading capability and the number and difficulty of books they are choosing to read. Of course, information on pupil learning is of no significance if it is not acted upon. Bangert-Drowns et al (1991) concluded from their meta-analysis that feedback improved learning effectiveness only where its 'mindful reception' was assured. Teachers thus need to intervene appropriately with pupils (through informal reading counselling or other guidance), subsequently using the AR system to track the effect of their intervention.

AR thus facilitates frequent formative evaluation in an assess-teach-evaluate iterative cycle, enabling early teacher intervention. Fuchs and Fuchs (1986) conducted a meta-analysis of 21 controlled studies of curriculum-based measurement which was at

least bi-weekly, noting an average effect size on pupil achievement of 0.70. These authors commented that regular formative evaluation is inductive, empirical, ecologically valid, closely linked with daily educational decision-making, more reliable and valid than many predictive 'individual educational programmes', and thus more likely to maximise the effectiveness of instruction.

Other reviews have confirmed the utility of formative assessment (e.g. Crooks, 1988), emphasising the importance of quality as well as quantity of feedback. Black and Wiliam (1998) concluded that assessment that precisely indicated pupil strengths and weaknesses and provided frequent constructive individualised feedback led to significant learning gains, compared to traditional summative assessment. The active engagement of pupils in the assessment process was seen as critical, and self-assessment as an essential tool in self-improvement. Affective aspects, such as the motivation to respond to feedback and the belief that it made a difference, were also important.

Previous research evidence

There is a substantial quantity of research on the Accelerated Reader in the USA (see Topping, 1999, for an on-line review). Much of this indicates that use of AR is highly correlated with gains in reading test scores, but unequivocally interpreting singular causality is difficult. Of 12 studies of AR citing substantial outcome data, only one failed to find evidence of impact on achievement. However, these studies were of very mixed quality – many failed to control confounding variables and many lacked data on implementation integrity. In the literature there are also anecdotal accounts of overall school-library circulation figures increasing by as much as 500% after the introduction of AR (e.g. Poock, 1998), ameliorating any concern that pupils might merely switch reading activity to books for which AR quizzes were available.

A more substantial study was reported by Topping and Sanders (2000), who merged AR data on 62 739 pupils from Grades 2 to 8 in Tennessee schools with the Tennessee Value Added Assessment System (TVAAS) longitudinal teacher value-added effects database, and explored the relationships between these independently obtained measures. There was a consistently positive and statistically significant relationship between increased number of books read and value added in Grades 3 to 6. There was also a consistently positive and statistically significant relationship between percentage correct and value added, consistent across all grades, but only becoming positive at the level of 80% correct in Grades 3 and 4 and 85% in Grades 5 to 7. This was true at all levels of reading volume and pupil ability. However more than half the pupils were operating below this level, suggesting implementation integrity was very variable. This implied that some teachers were not attending or responding to AR At-Risk Reports. The differences between value-adding and nonvalue-adding teachers in relation to their effective management of reading volume and percentage correct was particularly striking in the lower grades, and especially for pupils of lower ability. Analysis of interactions confirmed that a higher volume of reading practice could yield higher reading achievement, but only when the reading practice was also characterised by a high percentage correct (i.e. was successful).

In the UK, although a handful of schools had used AR prior to the current study, only one previous study had been reported. This quasi-experimental evaluation (Vollands, Topping & Evans, 1999) explored the effects of AR on reading achievement and motivation in two schools in severely socio-economically disadvantaged areas. Results

suggested that the programme yielded gains in reading achievement for these at-risk readers which were superior to those from regular classroom teaching and an alternative intensive method. Additionally, the programme yielded significant gains in measured attitudes to reading for girls. The issue of implementation integrity was specifically addressed, the researchers noting it was initially poor in both experimental locations, improving over time, despite the experimental teachers receiving one day of training. In particular, less time was devoted to class silent reading practice in experimental than in comparison classes. The study thus suggested that AR was effective by improving the quality of engagement with literature by pupils, rather than merely increasing the quantity of reading practice (time on task at reading).

Aims of this study

The current study aimed to explore the impact of AR on achievement in reading in a larger number of schools in the UK. These schools were to be of different types and sizes and geographically spread throughout the country, thus constituting a wider sampling of UK schools. Gains in reading achievement were to be assessed by paper reading tests normed in the UK administered on a pre-post basis, and by a computer-based adaptive norm-referenced test normed in the USA. The latter was to be administered pre-post and also at an interim point to appraise developmental trends over time more sensitively. Variation between schools in implementation was expected, given their different contexts. The extent to which schools implemented AR well in their own context was explored by direct observation during researcher visits and by analysis of the implementation integrity data generated by the AR programme itself. Additionally, the use of two different types of reading assessment instrument for triangulation purposes enabled exploration of their co-validity and reliability.

Method

Sampling

Schools known to possess the AR software were invited to volunteer to participate. Schools were then selected to give some geographical spread, but also to include a disproportionately large number of schools in disadvantaged areas. Thirteen schools were finally identified. One was in a disadvantaged area in north-east Scotland, three in a disadvantaged area in north-east England, two each in the relatively advantaged areas of Cambridgeshire and Kent (southern England), two in a disadvantaged area of East London and three in a disadvantaged area in Croydon (south of London). Thus only four of the thirteen schools were not in disadvantaged areas.

Five were primary (elementary) schools, three were first schools (Grades K-4), one was a junior school (Grades 3–6), one was a middle school, three were high schools and one was a ‘City Technology College’ (a senior high school with extra funding for technology). Selection of two classes for participation within schools was driven by teacher interest and willingness to participate and the strategic inclinations of the school headteacher (principal). The participating pupils were between the ages of 7–14 years (Years/Grades 3 to 9). Sampling thus was opportunistic and involved considerable self-selection. Nevertheless, a broad range of schools and pupils participated.

Participant resources and training

Two members of staff from each of the pilot schools attended a one-day training seminar and a further later two-day training seminar prior to implementation. In the intervening period they had the opportunity to install, debug and try out AR in their classes. Telephone technical support was available throughout the project. The teachers were also given the evaluation instruments and relevant instructions for administration. All schools installed a minimum of 1000 book quizzes of their own selection to match their library resources. (Schools can order customised discs, and quizzes on English rather than North American classic books are available.)

Instruments

Reading achievement was assessed using two different group administered multiple-choice paper reading tests normed and widely used in the UK. These were the *Primary Reading Test* (France, 1981) and the *Group Reading Test II 6–14* (NFER-Nelson, 1998). The use of more than one paper test was intended to offer additional triangulation. Each paper test was available in two levels for pupils of different ages and also in two parallel forms at each level to avoid content practice effects (*Primary Reading Test* – Levels/Forms 1/1A and 2/2A, *Group Reading Test* – Levels/Forms A/B and C/D). The PRT covered the age-range 6–11 years, the GRT the age-range 6–14 years. Consequently, the GRT was used with all pupils in high or middle schools. In the other schools, where either PRT (Level 1 or 2) or GRT (Level A or C) could apply, PRT or GRT were allocated randomly, at the level appropriate to the age of the pupils. Tests were administered to the pupils by class teachers. The parallel form of the paper test was always used at post-test.

Gains in reading achievement in all schools were also assessed using a computer-based adaptive multiple-choice norm-referenced test (STAR Reading) (Advantage Learning Systems, 1997). STAR was normed in 1996 in a stratified sample of 42 000 pupils from 171 schools in 37 states across the USA. As STAR is a computerised adaptive test, evaluation of reliability through traditional split-half methods is not possible. Test-retest reliability for 34 446 pupils is cited as 0.85 to 0.95. Assaying validity against a wide range of other tests (18 in total) in Grades 1–4 yielded correlation coefficients ranging from 0.65–0.90 (mean for 29 comparisons with $n > 100 = 0.76$). In Grades 5–8 correlation coefficients ranged from 0.34–0.93 (mean for 30 comparisons with $n > 100 = 0.72$).

Variables

Both the UK paper tests expressed a pupil's test performance in terms of a standardised score and also a 'reading age'. The reading ages suffered from ceiling and (particularly) floor effects, lacking fine discrimination at low raw scores and high raw scores. These effects impacted different schools and classes to different degrees. The standardised scores did not suffer to nearly the same extent in this respect, and were regarded as the more sensitive indicator. The extent to which the pupils in the disadvantaged schools in this study might be expected to make 'normal' gains is open to debate. The test norms indicate 'normal gains' for a normal population, not for an atypically socio-economically disadvantaged school, in which pupils might usually make gains below the level which is 'normal' on a UK-wide basis. This has implications for the interpretation of outcome results in this study.

The US STAR computer tests expressed a pupil's test performance in terms of a 'grade equivalent' (GE). STAR decimal GE tenths run from $\times .0$ through $\times .9$ and relate directly to the months September to June, the months of July and August being regarded by default as not part of the grade (no growth assumed). Mapping STAR GE on to UK reading ages thus presents several calibration problems. Great caution is necessary in comparing North American and UK children on the basis of their Grade or Year level.

The variables paper standardised score, paper reading age and STAR grade equivalent seemed the most closely related between the UK and US tests, and were the basis of all subsequent comparisons. Within schools, analyses were conducted to explore any differences in outcomes between different genders of pupil, which might reflect differential impact by gender.

The AR programme itself automatically generates data that give some insight into implementation integrity. Previous studies (e.g. Topping and Sanders, 2000) have indicated the significance of mean percentage correct per pupil on AR quiz items as an important indicator of implementation integrity. The average percentage correct (arpc) per pupil was therefore entered into the analysis. Other AR indicators were:

- AR tests taken during period (artt);
- AR tests passed during period (artp);
- AR average points earned during period (arpe);
- AR average points possible during period (arpp);
- AR average reading level of books read during period (ararl).

The average number of tests taken and points earned within the AR programme over a period is an indicator of volume of successful reading and self-testing activity, while the average reading level gives some indication of the degree of challenge presented by the books chosen to the pupil.

The AR system also flags up those pupils considered 'at risk' – showing a pattern of AR reading data suggesting some dysfunctionality and a consequent need for some guidance or intervention by the teacher. At any point, the proportion of the class who are considered at risk should of course be quite small, since a large proportion suggests the teacher is failing to intervene in response to at-risk flagging in successive weeks. The software manufacturers suggest that the at-risk proportion should never be more than 10–15%. Therefore the proportion of pupils flagged as at-risk (arisk) was entered into the analysis. This variable is of course at the unit of analysis of the class rather than the individual pupil.

Timescale

Participating schools completed the pre-tests on the paper test and STAR in mid-September. They then implemented AR with the targeted experimental pupils. Schools completed the interim tests on STAR in mid-December, approximately 0.25 calendar years later, and continued to implement AR. Schools completed the post-tests on the paper test and STAR in mid-April, approximately 0.33 calendar years later. However, this period included two vacation periods each of two weeks, during which the pupils were not exposed to the programme, so AR activity during this period was constrained to 0.25 years, identical to the pre-interim period. The pre-post period was thus approximately 0.58 calendar years.

Analysis

As not all schools reliably returned data, some analyses remained incomplete. Consequently it was considered that glossing the data with complex statistical analysis was inappropriate. Participating pupils generally constituted the entirety of whole classes and thus could be considered normal for their context. Variances for within-group comparisons were generally very similar. Therefore, parametric analysis was deemed appropriate and comparison of means using Student's *t*-test for related samples the main form of analysis chosen. Comparisons were made only within groups, not between them (including those by gender). This was supplemented by parametric correlation analysis (Pearson's coefficient). Two-tailed tests of statistical significance were used throughout.

Results

Outcomes: aggregate analysis

The 13 schools included 23 experimental classes/teachers. The years/grades of the experimental classes were (n): 3 (4), 4 (3), 5 (5), 6 (5), 7 (1), 8 (4), 9 (1). The total number of experimental pupils was 704. The number of classes using each paper test were as follows: *Primary Reading Test* (Forms 1/1A) n = 2, *Primary Reading Test* (Forms 2/2A) n = 4, *Group Reading Test* (Forms A/B) n = 4, *Group Reading Test* (Forms C/D) n = 13.

On norm-referenced tests, the average growth of all pupils in reading performance or reading age in one calendar year can be taken as statistically normal, indicated by the standardised score remaining the same. On the locally normed paper test, on average pupils progressed in tested reading skills at greater than normal rates. The difference between pre- and post-test was highly statistically significant (see Table 1).

Table 1. Aggregate pre-post gains on paper reading tests.

	N	mean	SD	gain	significance
Standardised score					
Pre-test	559	96.93	15.45	2.84	$p<0.001$
Post-test	559	99.77	18.13		
Reading age (years)					
Pre-test	562	10.17	4.30	0.61	$p<0.001$
Post-test	562	10.78	2.74		

On the STAR test, pupils also progressed in tested reading skills at greater than normal rates, the gain again being highly statistically significant (see Table 2).

In each class, at pre- and interim-testing, the numbers of pupils tested with STAR was typically larger than the number tested with the paper test (which had to be administered at one time regardless of pupil absence). This highlighted the usefulness of the individually self-administered computer-adaptive STAR test in terms of maximising pupil participation in assessment.

In pre-post testing, boys gained twice as much as girls on the local paper test, but girls did slightly better on STAR (see Table 3).

Overall, any differential impact of AR by gender thus appears likely to favour boys if anything, although this varied from school to school as well as from test to test.

Table 2. Aggregate pre-interim-post gains on STAR reading test: grade equivalents.

	n	mean	SD	gain	significance
Pre-post					
Pre-test	294	3.26	1.73	0.75	$p<0.001$
Post-test	294	4.01	1.93		
Pre-interim					
Pre-test	507	3.96	2.31	0.44	$p<0.001$
Post-test	507	4.40	2.47		
Interim-post					
Pre-test	295	3.68	1.89	0.32	$p<0.001$
Post-test	295	4.00	1.89		

Table 3. Aggregate pre-post test gains by gender and test.

	n	mean	SD	gain	significance
Paper test standardised score					
Males					
Pre-test	293	96.57	15.68	3.82	$p<0.001$
Post-test	263	100.39	18.03		
Females					
Pre-test	263	97.22	15.27	1.99	$p=0.012$
Post-test	263	99.21	18.31		
STAR test GE					
Males					
Pre-test	154	3.30	1.94	0.70	$p<0.001$
Post-test	154	4.00	2.03		
Females					
Pre-test	139	3.19	1.44	0.80	$p<0.001$
Post-test	139	3.99	1.72		

Implementation integrity: aggregate analysis

As AR implementation data were variously available for different schools for the pre-interim, interim-post and pre-post periods, comparison of averages across periods must proceed with caution. Nevertheless, these are provided in Table 4.

For all periods, the mean percentage correct is considerably lower than that recommended by the software designers for optimal effectiveness (85% correct) and found to be associated with value added on longitudinal tests (Topping and Sanders, 2000). Additionally, in many of the schools the proportion of pupils flagged as at risk was typically high, suggesting that teacher intervention in response to at-risk reports was low, and raising questions about the implementation integrity of the programme.

Outcomes of good implementation: a case study

School F is a primary (elementary) school, involving two classes/teachers in the field trial, in Years/Grades 4 and 6. The paper test used was GRT C/D. Over both classes, the paper test pre-post gains were very large, in standardised score (6.39) and reading age (1.03 years). Star pre-post GE gains were similarly large (1.05). The parallel Year/Grade 4 class gains scores were 4.92, 0.98 and 1.05, while the Year/Grade 6 class gain scores

Table 4. AR implementation integrity data: pre-interim-post.

Variable	Pre-interim (n = 445)	Interim-post (n = 194)	Pre-post (n = 222)
Average reading level of books read (ararl)	2.93	3.39	3.22
Tests taken during period (artt)	13.82	12.45	31.85
Tests passed during period (artp)	12.07	9.25	31.85
Average percentage correct (arpc) per pupil	71.52	73.48	75.62
Average points possible during period (arpp)	18.18	39.26	37.01
Average points earned during period (arpe)	12.56	14.44	24.15

were 7.76, 1.08 and STAR results not available. Although both classes showed large gains, on the surface the older class appeared to have gained more than the other class, despite being a little less able in relation to their age at pre-test. Considering outcome differences by gender over both classes, males performed somewhat better on the paper test, but females did better on the STAR test.

Implementation integrity was reported to be very high by the visiting researcher. The classes involved in the pilot programme appeared very accustomed to reading and quizzing. The pupils were aware of their individual targets, including average book level targets. In general, AR seemed to be an integral part of the school culture, and was actually in use by ten classroom teachers, although only two classes were providing evaluation data for this study. In many schools, impressions gleaned from visiting observation were not always supported by the AR implementation data. However, substantial AR data were received from this school (see Table 5).

Table 5. Pre-interim AR implementation integrity data school F by class.

Variable	Year 4 (n = 26)	Year 6 (n = 25)
Average reading level of books read (ararl)	3.25	4.34
Tests taken during period (artt)	4.31	15.52
Tests passed during period (artp)	3.96	14.56
Average percentage correct (arpc) per pupil	85.85	82.71
Average points possible during period (arpp)	7.50	69.28
Average points earned during period (arpe)	6.29	59.54

Thus, for both classes, these implementation indicators were much more in line with recommended levels than was the average for all field-trial schools. However, proportions of pupils at risk for the Year/Grade 4 class were 42% for the pre-interim period, 46% for the interim-post period. Proportions at risk for the Year/Grade 6 class were 19% for the pre-interim period, none available for the interim-post period owing to computer malfunction. This suggests that in the Year/Grade 4 class, teacher intervention in response to at-risk reports was low, and lower than recommended for maximum effectiveness of the programme. In the Year/Grade 6 class, during the pre-interim period teacher intervention in response to at-risk reports was high, and almost equivalent to the level recommended for maximum effectiveness of the programme. This might well have been reflected in superior test gains.

Equivalence of paper and STAR tests

Considering all the data and all tests, paper test standardised scores were correlated with STAR GE scores at both pre-test and post-test. The correlation at pre-test was 0.647

(n = 545), at post-test 0.598 (n = 301). On all paper tests the pre-post correlation was 0.760 (n = 716), while STAR pre-interim correlation was 0.903 (n = 520) and STAR interim-post correlation was 0.861 (n = 304). This suggests that in practice in this data-gathering context, the test-retest reliability of the paper tests fell considerably short of that noted in their technical manuals (although conventional test-retest procedures would not involve so long an inter-test period). The STAR tests appeared relatively reliable.

Conclusions

As might be expected in a field trial with only distant telephone support of implementation after training, quality of implementation of the AR programme was very varied. Additionally, the data returned by participating schools varied in completeness. The US-designed STAR computer test of reading appeared more stable in some respects than the paper tests of reading which had been devised and normed in the UK. STAR had the additional advantage of not requiring all testees to be present at the same time for testing, and therefore tended to yield more complete data.

Nonetheless, some interesting findings emerged. On average across all the schools, gain on all norm-referenced outcome measures was in excess of 'normal' rates and highly statistically significant. This was despite the majority of schools being in socio-economically disadvantaged areas where even 'normal' gains might not have been expected. This suggests that the Accelerated Reader had a significant impact on reading achievement overall. However, even within these schools, implementation integrity varied a great deal, and positive direct observations from visiting researchers were in a number of cases contradicted by implementation data gathered by the AR programme itself. One case-study school which came near to implementing the programme in the recommended way as indicated by both direct observation and AR data showed particularly high gains on tests of reading achievement.

A computerised curriculum-based learning information system (LIS) for reading (AR) enables more frequent, detailed, consistent and stable assessment of pupil reading comprehension of 'real' books. It appears to be a potentially useful tool for monitoring and managing increases in both the quantity and quality of reading practice, with consequent increases in reading achievement. Nonetheless, the information generated by the LIS will constitute mere noise unless received mindfully by teacher, pupil or (preferably) both. The LIS feedback to the teacher should lead to appropriate, timely and consistent intervention with the pupil.

More research is needed on the pedagogical implications of the use of such programmes. The impact of AR on affective aspects of reading and pupil self-management of learning require further exploration, particularly with respect to generalisation beyond programme activities. However, these are difficult to measure. Nonetheless, a programme that generates its own implementation integrity data offers considerable research advantages.

Acknowledgement

This research was supported by the University of Dundee (research assistant and data analysis) and by Renaissance Learning (supplying STAR programme, training and technical support to participating schools).

References

Advantage Learning Systems (1993). *The Accelerated Reader* (Computer Programme). Wisconsin Rapids, WI: ALS. (www.renlearn.com, www.renlearn.co.uk).

Advantage Learning Systems (1997). *STAR: Standardized Test for Assessment of Reading*. Wisconsin Rapids, WI: ALS. (www.renlearn.com, www.renlearn.co.uk).

Allington, R. (1984). Content coverage and contextual reading in reading groups. *Journal of Reading Behavior*, 16, 85–96.

Anderson, R.C., Wilson, P.T. & Fielding, L.G. (1988). Growth in reading and how children spend their time outside of school. *Reading Research Quarterly*, 23, 285–303.

Bangert-Drowns, R.L., Kulik, C-L.C., Kulik, J.A. & Morgan, M. (1991). The instructional effects of feedback in test-like events. *Review of Educational Research*, 61, 213–238.

Biemiller, A. (1978). Relationships between oral reading rates for letters, words and simple text in the development of reading achievement. *Reading Research Quarterly*, 2, 223–253.

Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5, 7–74.

Butler, D.L. & Winne, P.H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 65, 245–281.

Crooks, T.J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, 58, 438–481.

Donahue, P.L., Voelkl, K.E., Campbell, J.R. & Mazzeo, J. (1999). *NAEP 1998 reading report card for the nation and the states: March 1999*. [Online], Available: http://nces.ed.gov/nationsreportcard/pubs/main1998/ [March 15].

France, N. (1981). *The Primary Reading Test (Levels 1 & 2)*. Windsor: NFER-Nelson.

Fuchs, L.S. & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children*, 53, 199–208.

Kluger, A.N. & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119, 254–284.

Leinhardt, G. (1985). Instructional time: A winged chariot? In C. Fisher & D. Berliner (Eds.), *Perspectives on instructional time*. New York: Longman.

Lhyle, K.G. & Kulhavy, R.W. (1987). Feedback processing and error correction. *Journal of Educational Psychology*, 79, 320–322.

Manning-Dowd, A. (1985). *The effectiveness of SSR: A review of the research*. (Educational Resources Information Center Document Reproduction Service No. ED 276 970).

NFER-Nelson Publishing Company (1998). *Group Reading Test II 6–14* (3rd Edn). Windsor: NFER-Nelson.

Organisation for Economic Co-operation and Development (2002). *Reading for Change: Performance and Engagement across Countries: Results from PISA 2000*. Paris: OECD.

Poock, M.M. (1998). The Accelerated Reader: An analysis of the software's strengths and weaknesses and how it can be used to its best potential. *School Library Media Activities Monthly*, 14(9), 32–35 (Educational Resources Information Center No. EJ 565 465).

Rowe, K.J. (1991). The influence of reading activity at home on students' attitudes towards reading, classroom attentiveness and reading achievement: An application of structural equation modelling. *British Journal of Educational Psychology*, 61, 19–35.

Schunk, D.H. (1994). Self-regulation of self-efficacy and attributions in academic settings. In D.H. Schunk & B.J. Zimmerman (Eds.), *Self-regulation of learning and performance: Issues and educational applications* (pp. 75–99). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shany, M.T. & Biemiller, A. (1995). Assisted reading practice: Effects on performance for poor readers in Grades 3 and 4. *Reading Research Quarterly*, 30, 382–395.

Shapiro, E.S. & Cole, C.L. (1994). *Behavior change in the classroom: Self-management interventions*. New York & London: Guilford Press.

Snow, C.E., Burns, M.S. & Griffin, P. (Eds.) (1998). *Preventing reading difficulties in young children.* Report of the committee on the prevention of reading difficulties in young children, National Research Council. National Academy of Sciences. (Available: www.nap.edu/readingroom.enter2.cgi?030906418X.html)

Stanovich, K. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, 21, 360–407.

Taylor, B., Frye, B. & Maruyama, G. (1990). Time spent reading and reading growth. *American Educational Research Journal*, 27, 351–362.

Topping, K.J. (1999). Formative assessment of reading comprehension by computer: Advantages and disadvantages of the Accelerated Reader software. *Reading OnLine* (I.R.A.) [Online]. Available www.readingonline.org/critical/topping [November 4]. (hypermedia).

Topping, K.J. (2001). *Thinking reading writing: A practical guide to paired learning with peers, parents & volunteers*. New York & London: Continuum International.

Topping, K.J. & Paul, T.D. (1999). Computer-assisted assessment of practice at reading: A large scale survey using Accelerated Reader data. *Reading and Writing Quarterly*, 15, 213–231 (themed issue on Electronic Literacy).

Topping, K.J. & Sanders, W.L. (2000). Teacher effectiveness and computer assessment of reading: Relating value added and learning information system data. *School Effectiveness and School Improvement*, 11, 305–337.

Vollands, S.R., Topping, K.J. & Evans, H.M. (1999). Computerized self-assessment of reading comprehension with the Accelerated Reader: Action research. *Reading and Writing Quarterly*, 15, 197–211 (themed issue on Electronic Literacy).

Address for correspondence: Professor Keith Topping, Faculty of Education & Social Work, University of Dundee, Gardyne Road, Dundee DD5 1NY, Scotland, UK E-mail: *k.j.topping@dundee.ac.uk*