Accelerated Reader: U.K. Pilot 1999-2000

Summary Report

February 1, 2001

K. J. Topping & A. M. Fisher

Centre for Paired Learning
University of Dundee

www.dundee.ac.uk/psychology/kjtopping

Contents

WHAT IS ACCELERATED READER?

The Learning Information System known as Accelerated Reader® (AR™) (Advantage Learning Systems, 1993; now known as Renaissance Learning) enables free standing computer assisted assessment of comprehension of "real books". It is a curriculum based assessment, that summarises and analyses results to enable teachers to monitor both the quantity and quality of reading practice engaged in by the student. It is voluntarily self-administered by students, and specifically intended to have strong formative effects on subsequent learning.

Students using the programme read a book from over 39,000 titles on the AR list, then take a multiple-choice comprehension test on the book at the computer, which scores the test and keeps records. Each book has a maximum point value according to its length and difficulty. When the student self-tests, the computer awards points up to this maximum, according to their number of correct test responses. Students select their own preferred books and read at their own pace. Teachers may choose to allow students 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. The default condition is that students can test on a book only once.

The computer also provides the teacher with an automatically updated analysis of scores for individuals or whole classes, which indicates average percent correct, difficulty of books read, points earned, and other diagnostic information. Computer generated "Diagnostic" reports enable the teacher to guide each student's reading practice for maximum effectiveness.

Implemented well, Accelerated Reader features:

- more frequent assessment
- more detailed assessment
- in less time
- with greater consistency

- formative feedback to the student
- aims to raise meta-cognitive awareness
- aims to motivate students to read more, longer, harder books

- formative feedback to the teacher
- class-wide diagnostic information, including At-Risk alert
- helps the teacher promote & manage effective reading practice

The software originated in the US, has been localised for the UK, and is supported by staff development and in-service training opportunities, leading to a wider school development program known as "Reading Renaissance". AR is currently in over 50,000 schools worldwide and its use is spreading to other countries. The associated "Model Classroom Program" identifies and celebrates classrooms in which good practice in the implementation of Reading Renaissance has been evidenced.

PREVIOUS RESEARCH ON ACCELERATED READER

There is a substantial quantity of research on Accelerated Reader in the U.S. (see Topping, 1999; School Renaissance Institute, 2000). In the UK however, although a handful of schools in the UK had used AR, prior to the current study only one previous study had been reported.

This quasi-experimental action research evaluation (Vollands, Topping & Evans, 1999), explored the effects of AR on reading achievement and motivation in two schools in severely socio-economically disadvantaged areas. The 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 program 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 students, rather than merely increasing the quantity of reading practice (time on task at reading). It also suggested that AR was effective irrespective of the availability of extrinsic tangible reinforcement, which elicited virtually no interest in the participant students.

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 U.K.

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 US (STAR Reading®, or "STAR"), the latter being administered pre-post and also at an interim point to appraise developmental trends over time more sensitively.

Although AR is also intended to impact affective and meta-cognitive components of the reading process, these were not directly explored in this study.

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.

The implementation integrity data were then related to the reading achievement data.

The use of two different types of reading assessment instrument for triangulation purposes also enabled exploration of their co-validity and reliability.

METHOD

Sampling

Schools were selected to give some geographical spread. It was also intended to include a disproportionately large number of schools in disadvantaged areas. However, as schools were selectively invited to volunteer on the basis of being previously known to the then Advantage Learning Systems UK director, they can hardly be considered an unbiased sample.

Thirteen pilot schools willing to take part in the pilot were identified. One was in a disadvantaged area of north east Scotland, three in a disadvantaged area in north east England (in the Newcastle on Tyne Education Action Zone), two in the relatively advantaged area of Cambridgeshire (eastern England), two in the relatively advantaged area of Kent (southern England), two in a disadvantaged area of East London, and three in a disadvantaged Education Action Zone at Croydon (south of London). Thus only four of the thirteen schools were not in disadvantaged areas. Although not fully regionally representative, the weighting to the south of England maximised accessibility for training purposes.

Six were Primary (elementary) schools (schools A, B, C, F, G, H), two were "First" schools (grades K-4) (schools I, K), one was a "Junior" school (grades 3-6) (school D), one was a Middle school (school J), two were High schools (schools E, M), and one was a "City Technology College" (a senior high school with extra funding for technology) (school L).

Selection of classes within schools for participation was also driven by interest and willingness to participate, as well as the strategic inclinations of the school headteacher (Principal). Initially two classes in each school were targeted to participate, but this did not necessarily stay the same. The participating students were between the ages of 7-14 years (Grades/Years 3 to 9).

Sampling thus was thus opportunistic and involved considerable self-selection. Nevertheless, a broad range of schools and students participated.

Participant Training

Participant schools were provided with the software free of charge in exchange for agreement to provide data for the evaluation.

All schools received the full AR software engine and were given the opportunity to select 1000 quizzes. Most of these were from the US catalogue. (There was some concern about differences in "classic" libraries between countries, but schools were able to order customised discs, and over 1600 tests on "English" classics are currently available). Schools were able to continue to order from this bank of quizzes during the course of the year. Each school was also provided with 350 free books and the accompanying AR quizzes.

Training was provided free of charge (although schools had to meet the cost of supply teacher cover). Two members of staff from each of the 13 pilot schools attended a one-day training seminar in London in March 1999 and a further two-day training seminar on 13th and 14th September 1999. In the intervening period they had the opportunity to install and begin to try out AR and the STAR (see below) programmes in their schools. At the September meeting the teachers were given the paper tests and relevant instructions for administration.

Instrumentation

Gains in reading achievement were to be assessed using two different group administered 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 students 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 basic forms (1, 2) of the Primary Reading Test (France, 1981) were standardised in 1978 on 12,000 students in randomly selected schools throughout the UK, and the parallel forms (1A, 2A) were standardised in 1980-81 on 2,000 students. Reliability is cited as varying between 0.932 and 0.969 (Kuder-Richardson 20). Correlation between the alternative forms cited as 0.87. Co-validity with the Richmond Tests of Basic Skills is cited as 0.82, with the Holborn reading test as 0.85, with the Southgate reading test as 0.91, and with teachers' independent assessments as 0.82 – 0.89.

The Group Reading Test II 6-14 (NFER-Nelson, 1998) was standardised in 1996 on 3524 students (Forms A/B) and 2315 students (Forms C/D) in randomly selected schools throughout the UK. Reliability (internal consistency) is cited as 0.94 (Forms A/B) and 0.84-0.88 (Forms C/D) (K-R 21). Alternate form reliability is cited as 0.95 (Forms A/B) and 0.89 (Forms C/D). Very various validity is cited for Forms A/B – 0.70-0.86 with the Graded Word Reading Test, 0.75-0.85 with the Schonell Word Reading Test. For Forms C/D, validity information is given only in terms of alignment with independent teacher assessment, which is claimed to be high, but unwieldy to summarise here.

The PRT covered the age range 6-11 years, the GRT the age range 6-14 years. Consequently, the GRT was used with all students 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 students. The parallel form of the paper test was always used at post-test.

Gains in reading achievement were also to be assessed using a computer-based adaptive norm-referenced test (STAR) (Advantage Learning Systems, 1997), which was normed in the US.

STAR was normed in 1996 in a stratified sample of 42,000 students from 171 schools in 37 states across the US. As STAR is a computerised adaptive test, evaluation of reliability through traditional split-half methods is not possible. Test-retest reliability for 34,446 students is cited as 0.85 to 0.95. Assaying validity against a wide range of other tests (18 in total) in Grades 1 through 4 yielded correlation coefficients ranging from 0.65 – 0.90 (mean for 29 comparisons with n>100 = 0.76). In Grades 5 through 8 correlation coefficients ranged from 0.34 – 0.93 (mean for 30 comparisons with n>100 = 0.72).

All the tests were similar in structure, offering the testee progressively more difficult sentences with a key word omitted, which the testee was required to select from a set of multiple-choice options. However, STAR offered three or four answer choices, while the paper tests offered five. Individual STAR items have a time limit, although in practice very few students time out and then only on a few items, so STAR is not a highly time sensitive test. The paper tests had no time limit for individual items or the overall test (although practical time constraints might have arisen in some classrooms during administration). STAR is individually computer-administered, while the paper tests were done individually but in a group setting in each class. STAR requires very simple keyboard skills, the paper tests very simple pencil skills. STAR has US cultural content and spelling, the paper tests have UK cultural content and spelling.

Variables

Both the UK paper tests expressed a student's test performance in terms of a "standardised score" (which relates the student's performance to the normal distribution of all the students of the same chronological age in the norming sample, with a mean of 100 and a standard deviation of 15) and also a "reading age" (the chronological age at which an average student would perform at that level). The reading ages suffered from ceiling and (particularly) floor effects, lacking fine discrimination below low raw scores and above high raw scores. These effects impacted different schools and classes to different degrees. However, the standardised scores did not suffer to nearly the same extent in this respect, and should be regarded as the more sensitive indicator. Both reading ages and standardised scores were entered into the analysis.

The extent to which the students 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 students might usually make gains below the level which is "normal" on a UK-wide basis. This has implications for the interpretation of both Reading Ages and Standardised Scores in this study.

The US STAR computer tests expressed a student's test performance (inter alia) in terms of a "grade equivalent" (GE) (the normal US grade placement of students for whom the particular performance is typical or average, in relation to the US norming population).

STAR decimal GE tenths run from x.0 through x.9 and relate directly to the months September through June, the months of July and August being regarded by default

as not part of the grade (no growth assumed). GE thus does not relate to the actual chronological age of the child, but to the number of months from September through June which the child has progressed through the grade. Of course, different students will have started the grade level at different actual chronological ages.

Mapping STAR GE on to UK reading ages thus presents several calibration problems. In most US school districts, children may commence first grade at the age of 5 providing they will be 6 by September 1 in the school year of entry. Thus, typical first graders at the beginning of the school year (grade placement = 1.0) will usually be between 6.0 and 6.9 years old and will probably average 6.5 years old.

This suggests that the least worst continuous algorithm for US Grade – UK Year equivalence is: Chronological Age – 5.5 = Grade Level. Unfortunately reality is more complex than this.

In England, the law does not require a child to start school until the start of term following the child's fifth birthday. Compulsory school age is determined by dates set by the Secretary of State for the start of the Autumn, Spring and Summer terms - these are 31 August, 31 December and 31 March.

However, there is a statutory year of schooling **before** Year 1 (often termed the "reception" year).

Despite the legal minimum requirement, children typically start primary school at the age of 4, provided they will be 5 before 31 December in the school year of entry ("rising 5"). Thus, typical "reception" year children at the beginning of the school year will usually be between 4.66 and 5.66 years old and will probably average 5.16 years old. In fact, as local authorities (school districts) have considerable discretion in their local admissions policy beyond the statutory minimum, some children might be even younger than this.

Consequently, children typically start Year 1 at the age of 5.66 - provided they will be 6 before Christmas in the school year of entry. Thus, typical Year 1 children at the beginning of the school year will usually be between 5.66 and 6.66 years old and will probably average 6.16 years old.

English "Years" thus map onto North American "Grades" (Grade 1 = Year 1, etc) only very approximately). It would be more correct to propose two least worst continuous algorithms for US Grade – UK Year equivalence:

- U.S. Chronological age - 5.50 = Grade Level
- English Chronological age - 5.16 = Grade Level

(Of course, there is still the problem of the GE being based on tenths of the US school year, and not one tenth of the calendar year)

However, such a mapping ignores the fact that English children will have experienced a full year of statutory fulltime schooling before Year/Grade 1. While North American children might have experienced Kindergarten before Grade 1, English children might equally have experienced "Nursery School" before their first year of statutory schooling which is before their Year 1.

In Scotland the system and the law is different – entry to school is only yearly (not termly). The legal minimum requirement is that children must commence school at the start of the school year (typically August 15th) after they become 5 years of age. However, children are entitled to enter school so long as they will attain the age of 5 by the February 28 after admission in the previous August, and exercising this option is more typical (although again local authorities have considerable discretion to extend this downwards). Thus the typical child starting school would be 5.0 years at that point, and the typical child starting the second year of statutory schooling would be 6.0 years.

The Scottish numbering is more rational than in England, with "Primary 2" or (P2) being equivalent to the English Year 1, and "Primary 1" or (P1) being equivalent to the English un-numbered "reception" Year of statutory schooling before Year 1. Consequently, Scottish P numbers are very approximately equivalent to Grade + 1.

Thus it would be more correct to propose three least worst continuous algorithms for US Grade – UK Year equivalence:

- U.S. Chronological age - 5.50 = Grade Level
- English Chronological age - 5.16 = Grade Level
- Scottish Chronological age - 5.00 = Grade Level

(Of course, there is still the problem of the GE being based on tenths of the US school year, and not one tenth of the calendar year)

Great caution is therefore necessary in comparing North American and UK children on the basis of their Grade or Year level.

Nevertheless, the variables Paper Test Standardised Score, Paper Test 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.

However, stable and accurate STAR results are dependent upon the teacher inserting the correct grade level for all students in the software before testing commences. STAR takes the grade level registered by the teacher for the students and automatically assigns a decimal part of the year depending on the time in the school year of testing (while ignoring the existence of July and August for this purpose). However, a number of schools inserted the incorrect grade level in STAR at pre-test, and some failed to correct this at subsequent testings. Consequently, their STAR results had to be adjusted manually using the procedure outlined in the technical manual for this purpose.

In schools where a control or comparison group of students who had not had access to AR had also been tested, analyses were also conducted to explore any differences in outcomes between experimental and control groups. Unfortunately, some schools which had promised testing of control or comparison groups subsequently failed to deliver this.

Within schools, analyses were also conducted to explore any differences in outcomes between different classes taught by different teachers, which might reflect differences in implementation integrity between teachers, for instance.

Within schools and/or classes, analyses were also conducted to explore any differences in outcomes between different genders of student, which might reflect differential impact by gender.

Additionally, the AR program itself automatically generates data which give some insight into implementation integrity.

For example, previous studies (for example, Topping and Sanders, 2000) have indicated the significance of mean percentage correct per student on AR quiz items as an important indicator of implementation integrity, correlating highly with value added in terms of achievement gains. The average percent correct (arpc) per student was therefore entered into the analysis.

Other AR indicators entered into the analysis 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 student.

The AR system also flags up to the teacher those students whose performance on the AR system seems sufficiently dysfunctional for them to be considered "At Risk" – needing 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 designers suggest that the At Risk proportion should never be more than 10-15%. Therefore the proportion of students flagged as At Risk (arisk) was entered into the analysis. This variable was of course at the unit of analysis of the class rather than the individual student. Students flagged as At Risk in only one respect were dealt with as if equivalent to students flagged as At Risk in more than one respect.

However, some schools did not return AR data, some schools returned AR data only for the latter part of the experimental period, some schools returned AR data in the wrong (unusable) form, and some schools returned AR data wrongly dated (rendering it unusable).

Only eight schools returned substantial AR data. Three further schools returned a little AR data, relating only to the end of the experimental period – this was not analysable in detail. The remaining two schools returned no AR data.

Of the eight schools returning substantial AR data, some had cumulated AR data over the whole experimental period, but some had reset the cumulation midway through the experimental period. Consequently, only four schools provided AR data cumulated to the end of the pre-interim period and continuingly cumulated to the end

of the interim-post period, while four provided AR data cumulated to the end of the pre-interim period and AR data newly cumulated from the beginning of the interim-post period to the end of that period. This obviously affected how the AR indicators of implementation integrity could be related to outcome indicators.

AR data cumulated to the end of the pre-interim period and continuingly cumulated to the end of the interim-post period enabled comparison to the paper test pre-post changes in standardised scores (pfss) and the STAR pre-post changes in grade equivalent (pfstarge). AR data cumulated only to the end of the pre-interim period enabled comparison only to the STAR pre-interim changes in grade equivalent (pistarge). AR data newly cumulated from the beginning of the interim-post period to the end of that period enabled comparison only to the STAR interim-post changes in grade equivalent (ifstarge).

In the case of the At Risk data, for consistency the percentage of the class At Risk was calculated for all classes for both the pre-interim and interim-post period.

Time Scale

Participating schools completed the pre-tests on the paper test and STAR in mid-September 1999.

They then implemented AR with the targeted experimental students.

Participating schools completed the interim tests on STAR in mid-December 1999, approximately 3 months, or 13 weeks, 0.25 calendar years later.

They continued to implement AR.

Participating schools completed the post-tests on the paper test and STAR in mid-April 2000, approximately 4 months, or 17 weeks, or 0.33 calendar years later. However, this period included two vacation periods each of two weeks, during which the students were not exposed to the programme, so activity during this period was constrained to 13 weeks or 0.25 years, identical to the pre-interim period.

The pre-post period was thus approximately 7 months, 30 weeks, or 0.58 calendar years.

Analysis

Not all schools reliably returned the agreed data, so some analyses remain incomplete. Paper tests were administered by class teachers, possibly in varying circumstances despite the clear and detailed instructions given. Consequently it was considered that glossing the data with complex statistical analysis was inappropriate.

Participating students generally constituted the entirety of whole classes and thus could be considered normal for their context. The level of measurement was in a ratio scale. Variances for within-group comparisons were generally very similar, allowing for the usual increase in variation in post-intervention scores. Therefore, parametric statistical analyses were deemed appropriate.

Comparison of means using Student's t-test for related samples was the main form of analysis chosen. Although it is accepted that multiple use of such tests raises issues about the possibility of type 1 error, the proportion of significant results was in fact very large. Comparisons were made only within groups, not between them (including those by class/teacher and by gender). This was supplemented by parametric correlation analysis (Pearson's coefficient) and by linear regression analysis.

Two-tailed tests of statistical significance were used throughout. Exact probabilities are cited up to three places of decimals. Probabilities smaller than p=0.001 are rendered as p<0.001.

OUTCOME RESULTS

Outcome: Overall Aggregate Analysis

schools: n = 13
number of experimental classes/teachers: 23
grade of experimental classes (n): 3 (4), 4 (3), 5 (5), 6 (5), 7 (1), 8 (4), 9 (1)
total number of cases: n=769
experimental students: n=704
control students: n=65
number of classes using each paper test:
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

Outcome Differences by Experimental/Control Conditions:

For experimental group:

- Paper test Pre-Post Standardised Score (pre: n=559, m=96.93, s.d.=15.45; post: n=559, m=99.77, s.d.=18.13; p<0.001, **significant gain**, 2.84)
- Paper test Pre-Post Reading Age (pre: n=562, m=10.17, s.d.=4.30; post: n=562, m=10.78, s.d.= 2.74, p<0.001, **significant gain**, 0.61)
- STAR Grade Equivalent Pre-Post (pre: n=294, m=3.26, s.d.=1.73; post: n=294, m=4.01, s.d.=1.93; p<0.001, **significant gain**, 0.75)
- STAR GE Pre-Interim (pre: n=507, m=3.96, s.d.=2.31; post: n=507, m=4.40, s.d.=2.47; p<0.001, **significant gain**, 0.44)
- STAR GE Interim-Post (pre: n=295, m=3.68, s.d.=1.89; post: n=295, m=4.00, s.d.=1.89; p<0.001, **significant gain**, 0.32)

STAR data at post-test were returned for many fewer students, with implications for interpretation. Leaving this aside, there is some evidence from STAR of a tendency for larger gains to be made pre-interim than interim-post (0.44 GE in 0.25 years pre-interim = 176% growth in relation to "normal" rates of gain; 0.32 GE in 0.33 calendar years interim-post = 97%). However, to some extent this difference is artefactual, as the interim-post period contained two vacation periods each of two weeks, during

which the students were not exposed to the programme. Considering the interim-post period of activity only (0.25 years), the growth during this period could be construed to be 128%.

Considering only the cases for which STAR data at post-test were available, the findings were similar: mean gain pre-interim = 0.42; mean gain interim-post = 0.31.

Considering only the cases for which STAR data at post-test were not available, the findings were similar: mean gain pre-interim = 0.44.

The control/comparison group data are minimal and derived from only two schools, and cannot validly be compared to the aggregated experimental student data, but are included here for completeness.

For control/comparison group:

- Paper test Pre-Post Standardised Score (pre: n=26, m=100.00, s.d.=13.31; post: n=26, m=103.38, s.d.=18.48; p=0.369, not significant)
- Paper test Pre-Post Reading Age (pre: n=26, m=12.85, s.d.=2.06; post: n=26, m=13.44, s.d.=2.16; p=0.223, not significant)
- STAR Grade Equivalent Pre-Post (pre: n=10, m=6.79, s.d.=1.22; post: n=10, m=6.67, s.d.=1.55; p=0.852 not significant)
- STAR GE Pre-Interim (pre: n=13, m=6.26, s.d.=1.77; post: n=13, m=6.19, s.d.=1.44; p=0.831, not significant)
- STAR GE Interim-Post – data not available

Outcome Differences by Gender:

For males:

- Paper test Pre-Post Standardised Score (pre: n=293, m=96.57, s.d.=15.68; post: n=293, m=100.39, s.d.=18.03; p<0.001, **significant gain**, 3.82)
- STAR Grade Equivalent Pre-Post (pre: n=154, m=3.30, s.d.=1.94; post: n=154, m=4.00, s.d.=2.03; p<0.001, **significant gain**, 0.70)

For females:

- Paper test Pre-Post Standardised Score (pre: n=263, m=97.22, s.d.=15.27; post: n=263, m=99.21, s.d.=18.31; p=0.012, **significant gain**, 1.99)
- STAR Grade Equivalent Pre-Post (pre: n=139, m=3.19, s.d.=1.44; post: n=139, m=3.99, s.d.=1.72; p<0.001, **significant gain**, 0.80)

Implementation Integrity

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 extreme caution.

For the pre-interim period (n=445), AR average reading level of books read during period (ararl) = 2.93, AR tests taken during period (artt) = 13.82, AR tests passed during period (artp) = 12.07, Average percent correct (arpc) per student = 71.52, AR average points possible during period (arpp) = 18.18, and AR average points earned during period (arpe) = 12.56.

For the interim-post period (n=194), AR average reading level of books read during period (ararl) = 3.39, AR tests taken during period (artt) = 12.45, AR tests passed during period (artp) = 9.25, Average percent correct (arpc) per student = 73.48, AR average points possible during period (arpp) = 39.26, and AR average points earned during period (arpe) = 14.44.

Thus from the pre-interim to interim-post periods, tests taken and passed and points earned seem to reduce slightly - this might be associated with students seeking to read harder books in the later period - although percentage correct increased slightly.

For the pre-post period (n=222), AR average reading level of books read during period (ararl) = 3.22, AR tests taken during period (artt) = 31.85, AR tests passed during period (artp) = 27.35, Average percent correct (arpc) per student = 75.62, AR average points possible during period (arpp) = 37.01, and AR average points earned during period (arpe) = 24.15.

For all periods, the mean percentage correct is considerably lower than that recommended by the software designers (85 per cent correct) and found to be associated with value added on longitudinal tests (Topping and Sanders, 2000).

In all but three (F, I, K) of the eight schools which provided substantial AR data, the proportion of students flagged as At Risk were typically high or very high, for both the pre-interim and interim-post period, and certainly much higher than the proportion recommended by the manufacturers. This suggests that teacher intervention in response to At Risk reports was very low, and much lower than recommended for maximum effectiveness of the programme. Further data from these three schools are given in the case study section below.

In the four schools which provided substantial AR data cumulated over the whole experimental period (schools I, K, M, B), no statistically significant correlations were found between the AR implementation variables and the Paper test or STAR test outcome variables.

In the eight schools which provided substantial AR data cumulated over the pre-interim period, no statistically significant correlations were found between the AR implementation variables and the STAR test outcome variables.

In the four schools which provided substantial AR data cumulated over the interim-post period (schools C, F, G, H), no statistically significant correlations were found between the AR implementation variables and the STAR test outcome variables, other than a very modest correlation with AR tests passed.

Consequently, further analysis of the AR implementation variables and their relationship to the Paper test or STAR test outcome variables focused upon the three schools with lower At Risk proportions (F, I, K).

Summary:

On the locally normed paper test, students progressed in tested reading skills at greater than normal rates, this gain being highly statistically significant.

As many of the students resided in socio-economically deprived areas, without intervention they might be expected to make gains at lower rates than are "normal" for the whole population. Their achievement was raised from below average levels to average levels.

On the STAR test, students also progressed in tested reading skills at greater than normal rates, the gain being highly statistically significant.

On the STAR test, there was evidence of greater growth during the pre-interim period than during the interim-post period. However, to some extent this difference is artefactual. Considering the interim-post period of activity only (0.25 years), the growth during this period could be construed to be 128%. The remaining difference in growth between pre-interim and interim-post periods might reflect naturalistically declining developmental trends in student reading growth during the school year, and/or might reflect a short-term novelty effect of the intervention, or be attributable to other factors.

From pre-post testing, boys gained twice as much as girls on the local paper test, but girls did slightly better on STAR. However, as the STAR post-test data are based on many fewer cases than the pre-test STAR data, this finding might not be reliable. 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.

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

In the schools which provided substantial AR data, the mean percentage correct on AR tests was considerably below that recommended by the software designers for optimal effectiveness.

In most of the schools which provided substantial AR data, the proportion of students flagged as At Risk were typically high or very high, suggesting that teacher intervention in response to At Risk reports was very low, and raising questions about the implementation integrity of the programme.

The three schools in which evidence of good implementation was available both from direct observation by a visiting researcher and by computer-gathered AR implementation data are discussed in more detail below. Over all the participating schools, no statistically significant correlations were found between the AR implementation variables and the Paper test or STAR test outcome variables.

Outcome: Best Evidence Synthesis

Some schools implemented AR very poorly or not at all during the period under investigation. Others appeared to implement AR somewhat more adequately but provided very partial and/or very unstable data.

In order to conduct a "best evidence synthesis", all data from schools which had implemented AR very poorly or not at all were discarded. These were Schools D, E, J and L.

All these schools had target students in the upper age ranges, and all used the GRT C/D as the paper test. Although in some ways these schools could be regarded as forming a comparison group, because the age profile of their students was so different from that of the remaining students, valid comparisons were not possible.

All data which were considered unstable and completely implausible were also discarded. These included:

- School H: paper test data (PRT 2) for DG class only
- School I: paper test data (PRT 1) for AH class only
- School M: paper (GRT C/D) and STAR test data for classes HW and BW; and for class CI controls (n=3)

Aggregate Outcome Results from Best Evidence Schools/Data:

schools: n = 9
number of experimental classes/teachers: 16
grade of experimental classes (n): 3 (4), 4 (3), 5 (5), 6 (3), 8 (1)
total number of cases: n=459

- Paper test Pre-Post Standardised Score (pre: n=340, m=94.19, s.d.=14.17; post: n=340, m=98.65, s.d.=16.70; p<0.001, **significant gain**, 4.46)
- Paper test Pre-Post Reading Age (pre: n=340, m=9.05, s.d.=2.06; post: n=340, m=10.01, s.d.=2.38; p<0.001, **significant gain**, 0.96, 166% growth, probably an under-estimate owing to floor and ceiling effects)
- STAR Grade Equivalent Pre-Post (pre: n=267, m=3.17, s.d.=1.51; post: n=267, m=3.92, s.d.=1.66; p<0.001, **significant gain**, 0.75, 129% growth)
- STAR GE Pre-Interim (pre: n=413, m=3.36, s.d.=1.47; post: n=413, m=3.86, s.d.=1.67; p<0.001, **significant gain**, 0.50, 200% growth)
- STAR GE Interim-Post (pre: n=269, m=3.58, s.d.=1.53; post: n=269, m=3.89, s.d.=1.59; p<0.001, **significant gain**, 0.31, 94% growth)

Since most of the excluded poor implementing schools had not provided adequate STAR data anyway, especially at post-test, the best evidence STAR results are little different from the overall aggregated results.

However, the best evidence paper test results are considerably better than the overall aggregated paper test results.

Outcome Differences by Gender from Best Evidence Schools/Data:

For males:

- Paper test Pre-Post Standardised Score (pre: n=177, m=93.84, s.d.=14.78; post: n=177, m=99.19, s.d.=17.32; p<0.001, **significant gain**, 5.35)
- Paper test Pre-Post Age (pre: n=177, m=9.07, s.d.=2.15; post: n=177, m=10.11, s.d.=2.52; p<0.001, **significant gain**, 1.04, growth 179%)
- STAR Grade Equivalent Pre-Post (pre: n=140, m=3.16, s.d.=1.97; post: n=140, m=3.87, s.d.=1.77; p<0.001, **significant gain**, 0.71, growth 122%)

For females:

- Paper test Pre-Post Standardised Score (pre: n=160, m=94.69, s.d.=13.59; post: n=160, m=98.26, s.d.=16.09; p<0.001, **significant gain**, 3.57)
- Paper test Pre-Post Reading Age (pre: n=160, m=9.04, s.d.=1.99; post: n=160, m=9.92, s.d.=2.23; p<0.001, **significant gain**, 0.88, growth 152%)
- STAR Grade Equivalent Pre-Post (pre: n=127, m=3.18, s.d.=1.32; post: n=127, m=3.97, s.d.=1.53; p<0.001, **significant gain**, 0.79, growth 136%)

As with the overall aggregate analysis, boys gained more than girls on the paper test, from a slightly lower baseline, but now these differences were less pronounced.

As with the overall aggregate analysis, girls gained a little more than boys on the STAR test, from a very slightly higher baseline.

Summary:

In the nine schools of the 13 which had implemented AR at all adequately, gain on all outcome measures was large and statistically significant at well beyond the p=0.001 level.

The best evidence paper test results were considerably better than the overall aggregated paper test results.

There was again some evidence from the computer test that proportionately more of this gain tended to be made in the pre-interim period. As discussed above, this might be partially artefactual.

However, the gains on the paper tests were in any event larger than those on the computer test, so the gains on the latter are obviously in no way spurious.

Boys gained more than girls on the paper test, from a slightly lower baseline, while girls gain a little more than boys on the computer test, from a very slightly higher baseline. However, these differences were small. There was little evidence that girls were significantly disadvantaged by the use of computer tests.

CASE STUDIES

As the aggregate analysis subsumes considerable variation in implementation integrity and outcome results between schools, more detailed Case Studies of the three schools with lower At Risk proportions (F, I, K) are presented in this section.

In the ensuing section (School Summaries), a brief summary of the data from each of the remaining schools is presented.

School F

school type: primary (elementary) school
number of classes/teachers: 2
grade(s): 4 and 6
experimental/control: 2 experimental classes only
paper test used: GRT C/D

Overall Outcome Differences:

- Paper test Pre-Post Standardised Score (pre: n=48, m=102.40, s.d.=17.55; post: n=48, m=108.79, s.d.=21.48; p<0.001, **significant gain**, 6.39)
- Paper test Pre-Post Reading Age (pre: n=48, m=10.73, s.d.=2.45; post: n=48, m=11.76, s.d.=2.77; p<0.001, **significant gain**, 1.03)
- STAR Grade Equivalent Pre-Post (one class only) (pre: n=25, m=3.35, s.d.=1.77; post: n=25, m=4.40, s.d.=1.84; p<0.001, **significant gain**, 1.05)
- STAR GE Pre-Interim (pre: n=51, m=4.00, s.d.=1.89; post: n=51, m=4.57, s.d.=2.09; p<0.001, **significant gain**, 0.57)
- STAR GE Interim-Post (one class only) (pre: n=24, m=3.85, s.d.=1.84; post: n=24, m=4.51, s.d.=1.79; p=0.010, **significant gain**, 0.66)

Outcome Differences by Class/Teacher:

For class/teacher: AJ (Grade 4)

- Paper test Pre-Post Standardised Score (pre: n=23, m=106.43, s.d.=17.44; post: n=23, m=111.35, s.d.=20.15; p=0.012, **significant gain**, 4.92)
- Paper test Pre-Post Reading Age (pre: n=23, m=10.56, s.d.=2.33; post: n=23, m=11.54, s.d.=2.79; p=0.007, **significant gain**, 0.98)
- STAR Grade Equivalent Pre-Post (pre: n=25, m=3.35, s.d.=1.77; post: n=25, m=4.40, s.d.=1.84; p<0.001, **significant gain**, 1.05)
- STAR GE Pre-Interim (pre: n=25, m=3.45, s.d.=1.70; post: n=25, m=3.86, s.d.=1.80; p=0.070, not significant, gain 0.41)
- STAR GE Interim-Post (pre: n=24, m=3.85, s.d.=1.84; post: n=24, m=4.51, s.d.=1.79; p=0.010, **significant gain**, 0.66)

For class/teacher: VH (Grade 6)

- Paper test Pre-Post Standardised Score (pre: n=25, m=98.68, s.d.=17.15; post: n=25, m=106.44, s.d.=22.79; p=0.001, **significant gain**, 7.76)
- Paper test Pre-Post Reading Age (pre: n=25, m=10.88, s.d.=2.60; post: n=25, m=11.96, s.d.=2.80; p=0.004, **significant gain**, 1.08)

- STAR GE Pre-Interim (pre: n=26, m=4.53, s.d.=1.93; post: n=26, m=5.23, s.d.=2.16; p<0.001, **significant gain**, 0.07)

Outcome Differences by Gender:

For males:

- Paper test Pre-Post Standardised Score (pre: n=25, m=103.36, s.d.=19.21; post: n=25, m=110.68, s.d.=24.10; p=0.001, **significant gain**, 7.32)
- STAR Grade Equivalent Pre-Post (one class only) (pre: n=15, m=3.68, s.d.=2.12; post: n=15, m=4.57, s.d.=2.15; p<0.001, **significant gain**, 0.89)

For females:

- Paper test Pre-Post Standardised Score (pre: n=23, m=101.35, s.d.=15.90; post: n=23, m=106.74, s.d.=18.53; p=0.005, **significant gain**, 5.39)
- STAR Grade Equivalent Pre-Post (one class only) (pre: n=10, m=2.86, s.d.=0.923; post: n=10, m=4.14, s.d.=1.31; p=0.002, **significant gain**, 1.28)

Implementation Integrity:

Reported to be Very High by visiting researcher. The class involved in the pilot program 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 used by 10 classroom teachers.

Substantial AR data were received, although not cumulated over the whole pre-post period. For class AJ, for the pre-interim period (n=26), AR average reading level of books read during period (ararl) = 3.25, AR tests taken during period (artt) = 4.31, AR tests passed during period (artp) = 3.96, Average percent correct (arpc) per student = 85.85, AR average points possible during period (arpp) = 7.50, and AR average points earned during period (arpe) = 6.29.

For class VH, for the pre-interim period (n=25), AR average reading level of books read during period (ararl) = 4.34, AR tests taken during period (artt) = 15.52, AR tests passed during period (artp) = 14.56, Average percent correct (arpc) per student = 82.71, AR average points possible during period (arpp) = 69.28, and AR average points earned during period (arpe) = 59.54. Thus for both classes, the implementation indicators were much more in line with recommended levels than was the average for all schools.

Proportions At Risk for class AJ were 42% for the pre-interim period, 46% for the interim-post period. Proportions At Risk for class VH were 19% for the pre-interim period, none available for the interim-post period. This suggests that in class AJ teacher intervention in response to At Risk reports was low, and lower than recommended for maximum effectiveness of the programme. In class VH, this suggests that 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. However, no statistically significant correlations were found between the AR implementation variables and the STAR test outcome variables for either class, analysed separately.

Summary:

Implementation appeared Very High from direct observation, and this was supported by high Percentage Correct and low At Risk proportions. Very large gains were evident on both the local paper test (178% growth) and on STAR (181% growth). This growth was evenly spread over the pre-interim period (228% growth) and the interim-post period (200% growth). Both classes (grade 4 and grade 6) showed large gains, although the grade 6 class were a little less able in relation to their age at pre-test, but gained more than the other class. Boys tended to gain more than girls on the paper test, but girls more than boys on the STAR test.

School I

school type: first school (grades K-4)
number of classes/teachers: 2
grade(s): 3
experimental/control: 2 experimental classes only
paper test used: PRT 1/1A

A north-eastern England city school but not in the Education Action Zone.

Overall Outcome Differences:

There were such large and implausible differences between the two experimental classes on the paper test that overall aggregate outcomes were not calculated.

- STAR Grade Equivalent Pre-Post (pre: n=47, m=2.61, s.d.=0.93; post: n=47, m=3.30, s.d.=1.01; p<0.001, **significant gain**, 0.69)
- STAR GE Pre-Interim (pre: n=47, m=2.63, s.d.=0.90; post: n=47, m=2.93, s.d.=0.95; p=0.008, **significant gain**, 0.30)
- STAR GE Interim-Post (pre: n=47, m=2.94, s.d.=0.95; post: n=47, m=3.35, s.d.=0.94; p<0.001, **significant gain**, 0.41)

Outcome Differences by Class/Teacher:

For class/teacher: AH

The Paper Reading Age results in particular are implausible, and might reflect unusual testing conditions at post-test.

- Paper test Pre-Post Standardised Score (pre: n=23, m=108.30, s.d.=13.64; post: n=23, m=104.22, s.d.=12.63; p=0.082, not significant, decline 4.08)
- Paper test Pre-Post Reading Age (pre: n=23, m=12.14, s.d.=17.23; post: n=23, m=8.82, s.d.=1.39; p=0.372, not significant, decline 3.32)
- STAR Grade Equivalent Pre-Post (pre: n=23, m=2.86, s.d.=0.90; post: n=23, m=3.55, s.d.=1.06; p<0.001, **significant gain**, 0.69)
- STAR GE Pre-Interim (pre: n=22, m=2.94, s.d.=0.85; post: n=22, m=3.21, s.d.=0.99; p=0.165, not significant, gain 0.27)

- STAR GE Interim-Post (pre: n=23, m=3.22, s.d.=0.97; post: n=23, m=3.66, s.d.=0.89; p=0.009, **significant gain**, 0.44)

For class/teacher: DM

- Paper test Pre-Post Standardised Score (pre: n=23, m=97.39, s.d.=12.34; post: n=23, m=99.65, s.d.=15.60; p=0.111, not significant, gain 2.26)
- Paper test Pre-Post Reading Age (pre: n=23, m=7.45, s.d.=1.03; post: n=23, m=8.46, s.d.=1.35; p<0.001, **significant gain**, 1.01)
- STAR Grade Equivalent Pre-Post (pre: n=24, m=2.37, s.d.=0.902; post: n=24, m=3.06, s.d.=0.910; p<0.001, **significant gain**, 0.69)
- STAR GE Pre-Interim (pre: n=25, m=2.36, s.d.=0.88; post: n=25, m=2.68, s.d.=0.85; p=0.012, **significant gain**, 0.32)
- STAR GE Interim-Post (pre: n=24, m=2.67, s.d.=0.87; post: n=24, m=3.06, s.d.=0.91; p=0.003, **significant gain**, 0.39)

Outcome Differences by Gender:

As analysis by gender aggregated across classes, given the difference between the classes on the paper test, analysis on the latter was disregarded.

For males:

- STAR Grade Equivalent Pre-Post (pre: n=20, m=2.60, s.d.=1.04; post: n=20, m=3.11, s.d.=1.14; p=0.005, **significant gain**, 0.51)

For females:

- STAR Grade Equivalent Pre-Post (pre: n=27, m=2.62, s.d.=0.86; post: n=27, m=3.44, s.d.=0.89; p<0.001, **significant gain**, 0.82)

Implementation Integrity:

Reported to be Very High by the visiting researcher. Both pilot classrooms as well as two other classrooms were using the program very enthusiastically. Pupils were given in-class reading/quizzing time every day, and appeared to be enjoying the program. All four of the teachers using the program had been asked to give presentations to teachers at other schools in the LEA on how to use AR, as several local schools had recently purchased the latest version of the software.

Substantial AR data were received, cumulated over the whole pre-post period. For class AH, for the pre-post period (n=25), AR average reading level of books read during period (ararl) = 2.91, AR tests taken during period (artt) = 47.16, AR tests passed during period (artp) = 40.24, Average percent correct (arpc) per student = 81.42, AR average points possible during period (arpp) = 54.44, and AR average points earned during period (arpe) = 42.18.

For class DM, for the pre-post period (n=25), AR average reading level of books read during period (ararl) = 2.81, AR tests taken during period (artt) = 44.92, AR tests passed during period (artp) = 40.36, Average percent correct (arpc) per student = 84.01, AR average points possible during period (arpp) = 35.48, and AR average points earned during period (arpe) = 26.66. Thus the implementation indicators for

both classes were much more in line with recommended levels than was the average for all schools.

Proportions At Risk for class AH were 40% for the pre-interim period, 40% for the interim-post period. Proportions At Risk for class DM were 40% for the pre-interim period, 28% for the interim-post period. This suggests that in class AH teacher intervention in response to At Risk reports was low, and lower than recommended for maximum effectiveness of the programme. In class DM teacher intervention in response to At Risk reports was low in the pre-interim period, but during the interim-post period increased to much nearer the level recommended for maximum effectiveness of the programme. However, no statistically significant correlations were found between the AR implementation variables and the Paper test or STAR test outcome variables for either class, analysed separately.

Summary:

Implementation integrity appeared to be good, as indicated by direct observation and AR data. Overall growth on STAR was evenly spread over the pre-interim (120%) and interim-post (124%) periods, and was similar in both classes. In the AH class, the paper test results appeared unstable and implausible. In the DM class, the paper test evidenced 174% growth. Girls did better than boys on both the paper and STAR tests.

School K

school type: first school (grades K-4)
number of classes/teachers: 2
grade(s): 3
experimental/control: 2 experimental classes only
paper test used: PRT 2/2A

A north-eastern England city school which was in the Education Action Zone.

Overall Outcome Differences:

- Paper test Pre-Post Standardised Score (pre: n=55, m=95.04, s.d.=14.78; post: n=55, m=99.64, s.d.=15.68; p=0.003, **significant gain**, 4.6)
- Paper test Pre-Post Reading Age (pre: n=55, m=7.75, s.d.=1.38; post: n=55, m=8.84, s.d.=1.45; p<0.001, **significant gain**, 1.09)
- STAR Grade Equivalent Pre-Post (pre: n=60, m=2.61, s.d.=0.97; post: n=60, m=3.37, s.d=1.12; p<0.001, **significant gain**, 0.76)
- STAR GE Pre-Interim (pre: n=62, m=2.58, s.d.=0.98; post: n=62, m=3.07, s.d.=1.07; p<0.001, **significant gain**, 0.49)
- STAR GE Interim-Post (pre: n=60, m=3.11, s.d. 1.05; post: n=60, m=3.37, s.d.=1.12; p=0.005, **significant gain**, 0.26)

Outcome Differences by Class/Teacher:

For class/teacher: JW

- Paper test Pre-Post Standardised Score (pre: n=21, m=97.00, s.d.=14.98; post: n=21, m=106.05, s.d.=14.63; p=0.001, **significant gain**, 9.05)
- Paper test Pre-Post Reading Age (pre: n=21, m=7.60, s.d.=1.21; post: n=21, m=8.85, s.d.=1.41; p<0.001, **significant gain**, 1.25)
- STAR Grade Equivalent Pre-Post (pre: n=25, m=2.49, s.d.=0.96; post: n=25, m=3.66, s.d.=1.08; p<0.001, **significant gain**, 1.17)
- STAR GE Pre-Interim (pre: n=26, m=2.43, s.d.=0.99; post: n=26, m=2.93, s.d.=1.13; p<0.001, **significant gain**, 0.50)
- STAR GE Interim-Post (pre: n=25, m=3.01, s.d.=1.07; post: n=25, m=3.66, s.d.=1.08; p<0.001, **significant gain**, 0.65)

For class/teacher: Mr A

- Paper test Pre-Post Standardised Score (pre: n=34, m=93.82, s.d.=14.75; post: n=34, m=95.68, s.d,=15.17; p=0.310, not significant, gain 1.86)
- Paper test Pre-Post Reading Age (pre: n=34, m=7.85, s.d.=1.48; post: n=34, m=8.83, s.d.=1.49; p<0.001, **significant gain**, 0.98)
- STAR Grade Equivalent Pre-Post (pre: n=35, m=2.70, s.d.=0.98; post: n=35, m=3.16, s.d.=1.11; p<0.001, **significant gain**, 0.46)
- STAR GE Pre-Interim (pre: n=36, m=2.69, s.d.=0.97; post: n=36, m=3.18, s.d.=1.02; p<0.001, **significant gain**, 0.49)
- STAR GE Interim-Post (pre: n=35, m=3.18, s.d.=1.04; post: n=35, m=3.16, s.d.=1.11; p=0.839, not significant, decline 0.02)

Outcome Differences by Gender:

For males:

- Paper test Pre-Post Standardised Score (pre: n=32, m=95.19, s.d.=13.77; post: n=32, m=99.06, s.d.=11.82; p=0.030, **significant gain**, 3.87)
- STAR Grade Equivalent Pre-Post (pre: n=32, m=2.55, s.d.=0.83; post: n=32, m=3.37, s.d.=1.10; p<0.001, **significant gain**, 0.82)

For females:

- Paper test Pre-Post Standardised Score (pre: n=23, m=94.83, s.d.=16.40; post: n=23, m=100.43, s.d.=20.14; p=0.050, **significant gain**, 5.60)
- STAR Grade Equivalent Pre-Post (pre: n=23, m=2.71, s.d.=1.18; post: n=23, m=3.37, s.d.=1.16; p<0.001, **significant gain**, 0.66)

Implementation Integrity:

Reported to be High by the visiting researcher. There was great enthusiasm for reading in both classrooms involved in the pilot program. Students were given 20-25 minutes of in-class reading time each day, and had a lot of access to the computer for quizzing throughout the day. Both teachers were very pleased with the additional target-setting and tracking capabilities of the latest version of the software, as they already set individual targets for each half-term. The only negative observation was

a lack of installed quizzes for particular reading and interest levels. UK quizzes were subsequently installed, so this was not a long-term problem.

Substantial AR data were received, cumulated over the whole pre-post period. For class JW, for the pre-interim period (n=27), AR average reading level of books read during period (ararl) = 2.73, AR tests taken during period (artt) = 46.93, AR tests passed during period (artp) = 43.33, Average percent correct (arpc) per student = 84.56, AR average points possible during period (arpp) = 28.35, and AR average points earned during period (arpe) = 23.06.

For class MrA, for the pre-interim period (n=35), AR average reading level of books read during period (ararl) = 2.81, AR tests taken during period (artt) = 54.34, AR tests passed during period (artp) = 49.37, Average percent correct (arpc) per student = 83.16, AR average points possible during period (arpp) = 37.01, and AR average points earned during period (arpe) = 27.78. Thus the AR implementation indicators for both classes were much more in line with recommended levels than was the average for all schools.

Proportions At Risk for class JW were 28% for the pre-interim period, 27% for the interim-post period. Proportions At Risk for class Mr A were 15% for the pre-interim period, 24% for the interim-post period. This suggests that in both classes teacher intervention in response to At Risk reports was high, and near to that recommended for maximum effectiveness of the programme (especially in Mr A's pre-interim period).

However, for class JW no statistically significant positive correlations were found between the AR implementation variables and the Paper test or STAR test outcome variables. For class MrA no statistically significant positive correlations were found between the AR implementation variables and the Paper test outcome variables, but modest and statistically significant correlations were found between the AR implementation variables Tests Passed (artp) (0.359, p=0.037, n=34) and Points Earned (arpe) (0.392, p=0.022, n=34) and the STAR GE test outcome variable.

Summary:

Implementation integrity appeared to be good, as indicated by direct observation and AR data. Despite high socio-economic deprivation, the students made greater than "normal" gains on the pre-post local paper test (155% growth) and on the pre-post STAR test (131% growth). However, one class (JW) did better than the other on both the paper test (216% growth) and on STAR (202% growth). In this class the STAR gains were evident over both the pre-interim and interim-post periods, whereas in the other class the STAR gains in the interim-post period were negligible. Girls did a little better than boys on the paper test, boys a little better than girls on the STAR test.

SCHOOL SUMMARIES

School A

School type: primary (elementary) school. An inner-city school in the Docklands area of East London, with a very large ethnic minority (Bengali) population. The positive impressions regarding implementation from visiting were not supported by the limited AR system data, but these latter might be unreliable, given staff changes and technology difficulties. Nevertheless, despite high socio-economic deprivation and a very large proportion of students for whom English was an Additional Language, the students made substantially greater than "normal" gains on the pre-post local paper test (140% growth) and on the pre-interim STAR test (160% growth). Boys made twice as much progress as girls on the pre-post local paper test.

School B

School type: primary (elementary) school. An inner-city school in the Docklands area of East London (Tower Hamlets), with 80% ethnic minority students (mostly Bengali), for whom English is an Additional Language. Implementation integrity appeared superficially good, but the AR data indicated underlying deficiencies. Despite high socio-economic deprivation and a very large proportion of students for whom English was an Additional Language, the students made substantially greater than "normal" gains on the pre-post local paper test (124% growth) and on the pre-interim STAR test (128% growth). This overall finding disguises some disparity between classes. The MB class showed much larger gains (171% growth) than the GF class on the local paper test, and achieved large gains on the pre-interim STAR (216% growth). However, the other class showed very large pre-post gains on STAR (148% growth), which were not available for the MB class. Interestingly, the bulk of this growth was achieved in the later interim-post period, perhaps suggesting delay in full implementation in this class. Boys did a little better than girls on the local paper test, but girls considerably better than boys on the STAR test.

School C

School type: primary (elementary) school. In an area of disadvantage which is an Education Action Zone, surrounded by a prosperous area. Implementation integrity appeared good from direct observation, but the AR data indicated underlying deficiencies. Despite high socio-economic deprivation, the students made greater than "normal" gains on the pre-post local paper test (122% growth) and on the pre-post STAR test (109% growth). The MD class showed much larger gains than the KM class on the local paper test (162% growth), but the opposite was true on the STAR test (MD 74% growth, KM 148% growth). In both classes, much of the growth appeared to be made in the pre-interim period. Boys did much better than girls on the local paper test, but girls did somewhat better than boys on the STAR test.

School D

School type: junior school (grades 3-6). In an area of disadvantage which is an Education Action Zone, surrounded by a prosperous area. There were no data to

indicate that implementation integrity was other than Low. Students in both classes gained at normal rates, with no significant differences between males and females.

School E

School type: high school. In an area of disadvantage which is an Education Action Zone, surrounded by a prosperous area. The intended control group teacher left the school and only the paper pre-test of this group was done, so these control data were not usable. Implementation Very Low. Students in both classes gained at normal rates. Classes were very disparate in ability at pre-test, and the Reading Age scores were possibly affected by ceiling effects in one class. Overall outcomes were similar for both classes. There were no test differences between boys and girls.

School G

School type: primary (elementary) school. A rural primary school, one of only a few in the country exempted from following the national Literacy Hour recommendations, as its record is so good. Implementation integrity appeared superficially good from direct observation, but the AR data indicated underlying deficiencies. Nevertheless, very large gains were evident on both the local pre-post paper test (250% growth) and pre-interim STAR (332% growth). These were evident in both classes (grade 5 and 6), although the JL class did better (these students were of lesser ability in relation to their age at pre-test than the other class). There was some evidence in one of the two classes of a slowing in growth in the interim-post period. Boys performed a little better than girls on the paper test, but girls performed better than boys on the STAR test.

School H

School type: primary (elementary) school. One of the Scottish schools in the Vollands et al. (1999) study, in a disadvantaged part of a city in north-east Scotland. This school was not visited by the researcher for direct observation of implementation. AR implementation data suggested some variation between classes and over time. Implementation quality therefore remains some. On the local paper test, overall the students progressed at normal rates (although that might be a "good" outcome in an area of high socio-economic deprivation). On the STAR test they did better (121% growth). However, much of this growth occurred in the pre-interim period. One class (DW) had good results on the paper test (140% growth), while the other (DG) had poor results on the paper test but the better results on the STAR test (143% growth). This suggests the DG class paper post-test results are unreliable. Boys did better than girls on both the paper test and on STAR.

School J

School type: middle school. A disadvantaged inner-city school in an Education Action Zone. Implementation Very Low. Students progressed at normal rates on the paper test and at less than normal rates on STAR. Growth on STAR was better during the interim-post period than during the pre-interim period, perhaps suggesting delayed implementation. Girls did much better than boys on the paper test, and the same as boys on the paper test.

School L

School type: city technology college (senior high). A large City Technology College in rural/prosperous area. Although the school initially appeared enthusiastic, organised and with excellent technology resources and a librarian, subsequently implementation was judged to be Very Low by the visiting researcher, owing to the combined effects of network problems and the unexpected absence of the teacher for the pilot classes who was also the AR coordinator. The nominally experimental classes had NOT been taking AR quizzes. The teachers intended to implement AR properly the following year. Data were also incomplete and confused. Although the notional experimental class gained twice as much on the paper test as the control group, this cannot be plausibly attributed to AR. Boys gained twice as much as girls on the paper test.

School M

School type: high school. This school is a rural comprehensive. Data were incomplete and confused. Although implementation quality seemed high from direct observation, the AR data revealed considerable deficiencies in this regard. In this school, experimental and control students were present in all three classes. It is not known what possibilities for contamination there might have been. Both experimental and control students declined on the paper test, the experimental students more than the controls. However, on the STAR test, the experimental students showed 160% growth, while the control students declined. Substantial differences between classes were evident. On the paper test, class HW showed a large decline, class BW a smaller decline, and class CI (of less able students) a substantial gain (136% growth). Classes HW and BW showed very modest pre-interim STAR gains, but class CI showed 400% growth! On the paper test, girls showed a much bigger decline than boys.

CO-VALIDITY OF INSTRUMENTATION

In the elementary schools which constitute most of the Best Evidence Synthesis sub-sample, paper test was allocated randomly to schools. An analysis of outcomes by paper tests was conducted. This suggested that gains on the GRT test tended to be higher than gains on the PRT test (5.6 vs. 3.3 points of standardised score). Consequently, if just one or the other paper test had been used in all schools, the overall paper test results might have been somewhat larger or smaller. Because the different levels of the same paper test are applied to different grades, any analysis between levels confounds student age with test level. Similarly, any analysis by student age is problematic.

Inter-test Correlation Analysis

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). Considering only the Best Evidence Synthesis 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 0.738 (n=372), at post-test 0.717 (n=214). Considering only the Best Evidence Synthesis data and only the Paper GRT Level C test (which had the largest n), standardised scores were correlated with STAR GE scores at both pre-test and post-test. The correlation at pre-test was 0.800 (n=204), at post-test 0.735 (n=62).

Intra-test Correlation Analysis

On the overall aggregate data, 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 the technical manual (although conventional test-retest procedures would not involve so long an inter-test period). However, examining the Best Evidence Synthesis data, the all paper tests pre-post correlation was 0.840 (n=340), while the STAR pre-interim correlation was 0.848 (n=413) and the STAR interim-post correlation was 0.804 (n=269). The STAR results should be interpreted in this context, and no assumption made that the local paper tests are innately more reliable.

Predicting Reading Age from STAR Grade Equivalent:

Given that the Paper test Reading Age was a somewhat unstable metric owing to floor and ceiling effects, a regression analysis to predict Paper test Standardised Score was conducted on the best evidence synthesis data, yielding the following equations: pre SS = 7.037 (pre STAR GE) + 70.680; post SS = 7.319 (post STAR GE) + 70.222. Applied to the Best Evidence Synthesis means above: pre SS predicted = 92.99, actual 94.19; post SS predicted 98.91, actual 98.65. When predicting Standardised Score from the best evidence data, estimation by linear regression appears stable and leaves relatively little of the variance unaccountable. Given the intra-test correlation evidence for the relative reliability of STAR in this context (mentioned above), these equations offer useful algorithms for further development.

Summary:

Of the paper tests, the GRT yielded larger gains than the PRT. Over all schools, the co-validity of the Paper tests and STAR was quite low. This might partially reflect the US cultural content of the STAR test. However, it was considerably higher using only best evidence data (possibly collected in more orderly environments). The test-retest reliability of the paper tests (in parallel forms) appeared considerably less in this context than stated in the technical manual, and the test-retest reliability of STAR compared very favourably with it. Attempts to predict Paper test scores from STAR Grade Equivalent appeared to be stable and potentially useful when based on the Best Evidence Synthesis data and predicting Standardised Score.

CONCLUSIONS

In the nine schools of the 13 which had implemented AR at all adequately, gain on all norm-referenced outcome measures was large and statistically significant at well beyond the p=0.001 level.

Given that the outcome measures were norm-referenced tests deployed in parallel or multiple forms, which in general indicated growth at considerably greater than normal rates of gain, this suggests that the intervention Accelerated Reader had a significant impact 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 disquieting implementation data gathered by the AR programme itself.

Only three schools came near to implementing the programme in the recommended way as indicated by both direct observation and AR data, and all three schools showed excellent gains on tests of reading achievement.

However, some other schools in which a lesser quality of implementation was evident nevertheless showed substantial gains on tests of reading achievement, suggesting a degree of robustness in the programme in a variety of school contexts.

Nonetheless, schools with very low implementation integrity tended to show only normal rates of gain on test.

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.

REFERENCES

Advantage Learning Systems (1993). *Accelerated Reader* (computer program). Wisconsin Rapids, WI: ALS. (www.renlearn.com/ar/default.htm).

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

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

NFER-Nelson Publishing Company (1998) *Group Reading Test II 6-14* (third edition). Windsor: NFER-Nelson.

School Renaissance Institute (2000) *Research summary*. Madison, WI: School Renaissance Institute.

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. & 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*(3), 305-37.

Vollands, S. R., Topping, K. J., & Evans, H. M. (1999). Computerized self-assessment of reading comprehension with the Accelerated Reader: Action research of impact on reading achievement and attitude. *Reading and Writing Quarterly, 15*(3), 197-211 (themed issue on Electronic Literacy).