Computers & Education 116 (2018) 176–190

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Implementation fidelity in computerised assessment of book reading

Keith Topping*

University of Dundee, Scotland, United Kingdom

ARTICLE INFO

Article history:
Received 16 March 2017
Received in revised form 24 August 2017
Accepted 21 September 2017
Available online 23 September 2017

Keywords:
Evaluation methodologies
Gender studies
Improving classroom teaching
Pedagogical issues
Teaching and learning strategies

ABSTRACT

Measuring the implementation fidelity (IF) or integrity of interventions is extremely important, since without it a positive or negative outcome cannot be interpreted. However, IF is actually measured relatively rarely. Direct and indirect methods of measurement have been used in the past, but tend to over-emphasize teacher behaviour. This paper focuses on student behaviour collated through computers - an interesting alternative. It deals with the reading of real books and reading achievement, for which variables a very large amount of computerised data was available – on 852,295 students in 3243 schools. Reading achievement was measured pre-post with STAR Reading, a computerised item-banked adaptive norm-referenced test of reading comprehension. IF came from the Accelerated Reader (AR), which measures understanding of independent reading of real books the student has chosen by a quiz. Results showed higher IF was related to higher achievement. Neither IF nor reading achievement related to socio-economic status. Primary (elementary) schools had higher IF and achievement than secondary (high) schools. Females had higher IF and achievement than males. Students of higher reading ability implemented AR at a higher level, but did not gain in reading at a higher level. However, this computerised method of measuring IF with book reading showed limited reliability, no greater than methods emphasising teacher behaviour. In future, IF measures emphasising student response and those emphasising teacher behaviour need to be blended, although the latter will never generate the sample size of the former. This may be true of implementation fidelity in areas other than book reading.

In these evidence-based times, there is much emphasis on randomised controlled trials (RCTs) as the “gold standard” of good research. However, as Stockard (2010) among others points out, there are concerns about the external validity of such findings. There are a number of collections of such research intended to impact on practice, but only one (The What Works Clearinghouse - WWC) is sponsored by the US government. Extraordinarily, the WWC largely disregards the issue of IF, assuming that on average it “washes out” in the reviews they promote.

Another issue with RCTs is that by definition they allocate the intervention randomly. In education, this means to teachers who may or may not have the slightest interest in implementing the intervention. Wehby, Maggin, Johnson, and Symons (2010) studied the effect that teacher choice of intervention had on their level and quality of implementation. A total of 69 teachers (88% female; 68% general education, 32% special education) working with K-6 students participated. Implementing a

* School of Education, University of Dundee, Dundee DD1 4HN, United Kingdom.
E-mail address: k.j.topping@dundee.ac.uk.

https://doi.org/10.1016/j.compedu.2017.09.009
0360-1315/

preferred intervention was related to higher degrees of initial and sustained IF as well as greater numbers of actual implementers.

Clearly, there are issues here about implementation fidelity. But how might it be defined?

1. Definition of implementation fidelity

Implementation fidelity (or integrity) was initially defined as the degree to which an intervention or treatment was implemented as planned, intended, or originally designed. However, this only specified the behaviour of the interventionist, not that of the recipients of the intervention. By contrast, Schulte, Easton, and Parker (2009) included features related to the delivery of the intervention, how the intervention was received by the participants, and how the participants were able to use the learned skills in a natural environment. Of course, the question then arises of which of these many indices are most related to outcome (Durlak & DuPre, 2008).

Despite the importance of treatment fidelity, historically it has been frequently overlooked in research and practice. Since the emphasis has moved towards "evidence-based" interventions, measuring the quality of intervention has become an increasing preoccupation. Clearly, there is little point attempting to implement an evidence-based intervention and measure the outcomes if there is no parallel attempt to see whether the method has actually been implemented. As Carroll et al. (2007) express it, IF acts as a potential moderator of the relationship between interventions and their intended outcomes. Unless IF is assessed, in a circumstance of poor outcome we cannot know whether the program did not work or merely was not implemented properly, or both. Indeed, even in a circumstance of good outcome, we also cannot know whether the program actually worked and was responsible for the positive outcome.

Dane and Schneider (1998) and Schulte et al. (2009) among others espoused five elements in IF often found in the previous literature: adherence to an intervention, exposure or dose, quality of delivery, participant responsiveness and program differentiation (the extent to which key factors in effectiveness are identified). Measuring IF is not easy - researchers quickly found that it was both complex and expensive. Not all interventions clearly specified what the teacher had to do and in what order. Indeed, some of them had optional teacher behaviours, assuming that no two teachers would implement alike. Indirect attempts which simply asked teachers whether they had implemented well were often found not to correlate with outcomes. Direct attempts which used observational methods (to avoid teacher subjectivity) were expensive (and consequently usable only on a small scale) and could still suffer from observer effects – what the teacher did when observed might not have been typical of what they did when not observed. Another issue was whether any professional development prior to the intervention was one-off, or whether it was several sessions with time in-between for reflection and discussion with colleagues, or included ongoing coaching to shape teacher behaviour as the program was implemented. Teacher behaviour is the focus of much of the literature. Schulte et al.'s (2009) inclusion of participant responsiveness has been largely overlooked. There is also an issue about how often IF should be assessed, since many of the reports in the literature are of short-term interventions.

2. The current paper

This paper is set in the context of the effectiveness of book reading and emphasises student response rather than teacher behaviour. The focus on book reading is because this area generates the largest amount of data of any computerised assessment. A parallel assessment of mathematics is available, but the number of users and amount of data is considerably smaller. The paper compares and contrasts two different kinds of participant indicators of IF with growth in reading achievement. Of course other measures may be relevant, but this study deploys measures of student response to counterbalance the existing over-emphasis on teacher behaviour. In this paper both outcome and IF measures are completed locally but scored online centrally, and the results fed back locally, all by computer. This central scoring enables the collection of large samples of data. Both IF and outcome measures generate a number of variables directly from student responses.

3. Previous research on IF in reading

The present paper interrogates the literature on IF in book reading by exploring research on indirect measures (self-reports completed subjectively by teachers and head teachers) and direct measures (completed by observation, although still far from "objective" given possible observer effects). Curiously, studies appear to have done one or the other – there are very few studies which have directly compared the two. After this, a further section explores the literature (such as it is) on IF using the novel measures deployed here – computer-assisted item-banked adaptive reading outcome assessment and assessment of comprehension of real books after they have been read.

3.1. Methodology of the literature review

The Social Sciences Citation Index (SSCI) and the Educational Research Information Centre (ERIC) were searched from 1995 to date (the terms implementation/treatment fidelity/integrity had little currency prior to this date). Search terms were "book reading" AND "implementation fidelity" OR "implementation fidelity" OR treatment integrity" OR "treatment fidelity". The inclusion criterion specified relevance to the research questions (see below). Only 33 hits resulted from the first search of titles and abstracts. A further criterion of incorporating substantive data was implemented. On reading the full text, a number

of the papers still proved to be opinion pieces, reducing the items for the final literature review to 17. Seven of these were indirect studies, seven were direct studies, and three concerned the AR/STAR combination. As Sanetti and Fallon (2011) pointed out, varied assessment of IF can influence interpretation of implementation.

3.2. Indirect studies

Darrow (2010) analysed 17 measures of fidelity used by 13 curriculum interventions. Overall, the studies insufficiently measured the primary components of fidelity. In many cases, the measures were more effective at assessing general quality of instruction and less successful at evaluating fidelity of implementation.

Components of a two-year school-wide intervention were studied by Feldman, Feighan, Kirtcheva, and Heereen (2012), aimed at bolstering middle school teachers' use of literacy strategies to raise students' reading achievement. Although at post-test students of intervention teachers had significantly higher Iowa Test of Basic Skills (ITBS) scores than students of non-participating teachers, no evidence was found that any relationship existed between teachers' fidelity of implementation and students' performance on the ITBS.

Fedor (2013) had 132 K-3 classroom teachers in 20 schools complete the Teacher's Implementation of Scientifically Based Reading Instruction (TISBRI) survey to investigate school-wide implementation of scientifically based reading instruction. There was no correlation between level of implementation and Grade 3 reading achievement.

The Response to Intervention Implementation Scale for Reading (RTIS-R) was developed by Noltemeyer, Boone, and Sansosti (2014). Data were collected from 53 principals and school psychologists implementing Response to Intervention in 33 schools in a Midwestern state. The results suggested the instrument showed a positive relationship with achievement data.

Balu et al. (2015) examined the implementation of Response to Intervention in grade 1–3 reading in 146 schools in 13 states. Full implementation was reported by 86% of the schools. In Grades 2 and 3 there were no significant impacts of intervention on reading scores.

A survey was distributed to principals of all intermediate, middle schools, and junior highs in the state of Texas by Williams (2015), investigating the relationship of best practice strategies with schools' academic achievement in math and reading. Varied rates of implementation were reported. Implementation was weakest for students living in poverty, where implementation quality had the largest relationship with student achievement.

Sharp, Sanders, Noltemeyer, Hoffman, and Boone (2016) collected data from 64 principals and school psychologists at 43 elementary schools. Hierarchical linear regression was used to examine the degree to which IF predicted student reading assessment results, when controlling for school demographic variables. IF significantly predicted student reading outcomes.

In summary, two studies showed some relationship between indirect IF and student achievement, but four studies did not and one study found many so-called “implementation” measures were not in fact measuring implementation. It seems that in general indirect measures tend not to predict student achievement reliably.

3.3. Direct studies

McIntyre et al. (2005) examined the implementation of ten early reading models. Results illustrated variability in IF. High implementers had much support; a practical, clear model; extensive professional development; or a combination of these. There was great variability across teachers in terms of instruction, primary instructional activities, the texts used, and how teachers used time.

Students in grades 2–5 and their teachers were studied by Henninger (2010), in an urban elementary school in the southwestern United States. Classroom teachers were observed implementing reading interventions. Path analysis was conducted to explore the relationship between two factors of implementation (intervention complexity and acceptability), treatment fidelity (adherence to intervention protocol) and student outcomes (oral reading fluency scores). Results indicated an inverse relationship between intervention complexity and treatment fidelity - when complexity was low, treatment fidelity was high and reading fluency scores were high. A positive relationship was also found between intervention acceptability and treatment fidelity - when acceptability was high, treatment fidelity was high and reading fluency scores were high.

Benner, Stage, and Ralston (2011) examined the extent to which program adherence and quality of delivery enhanced or constrained the effects of a reading intervention for 281 middle school students experiencing reading difficulties. Students made significant improvements in their basic reading skills and passage comprehension, but with variations. Fidelity of implementation accounted for 22% and 18% of the variance in gains in basic reading skills and passage comprehension respectively. Two teacher behaviours (following the lesson format as designed and re-teaching lessons when needed) predicted student performance above and beyond other teacher actions.

A comprehensive set of materials and procedures for observing use of a structured reading program were developed by Begeny, Upright, Easton, Ehrenbock, and Tunstall (2013) and Begeny, Easton, Upright, Tunstall, and Ehrenbock (2014). They related direct observations to teacher self-report and permanent products generated by the intervention. They found that the observation and feedback procedures were effective in producing strong IF and this correlated highly with self-report. However, permanent products did not correspond well with IF.

Fogarty et al. (2014) examined the implementation and effects of a multicomponent reading comprehension intervention in sixth-to eighth-grade English language arts classes in three schools. Participants were 14 teachers and 859 students. Their

IF framework included adherence, quality, dosage, program differentiation, and student responsiveness. All teachers taught both experimental and control conditions. There was no difference between the intervention and control groups on standardized or researcher-developed measures. A Confirmatory Factor Analysis revealed a single fidelity factor composed of all the variables they studied. Fidelity was statistically significantly related to outcomes. The authors commented that these findings underscored the complexity of implementing multicomponent interventions and the importance of measuring multiple dimensions of IF.

A systematic observational study of middle school educators (Grades 6–8) in two states who provided reading interventions was reported by Ciullo et al. (2016). Intervention sessions were examined to investigate (a) the frequency and type of evidence-based strategies implemented for students with learning disabilities/reading difficulties, and (b) whether observed practices aligned with researcher recommendations. Evidence-based interventions including explicit instruction, cognitive strategy instruction, content enhancements, and independent practice opportunities were reported infrequently and not related to independent measures of achievement. Instructional differences across sites were demonstrated.

In summary, the evidence for directly measured IF impacting on achievement is limited. One study demonstrated that complexity and acceptability were mediating variables in the relationship with achievement. Additionally, some components of fidelity were more predictive than others. Structured observation and feedback procedures yielded impact on IF, but this did not correlate with achievement.

It is noteworthy that most of these direct and indirect studies focused on a short period of implementation. Few report IF over a longer period such as a year. Even fewer report IF indices available as a matter of course without additional effort as during the implementation of AR. Additionally, most of them focused on teacher behaviour. A few used reading tests to show student progress, but this is not really an indicator of student responsiveness. Begeny et al. (2014) compared teacher self-report and observational data. Only Fogarty et al.'s (2014) study included student responsiveness.

3.4. Previous research on IF with AR/STAR reading

AR offers a novel way of assessing IF, by taking indices of student responsiveness directly through computers (see Methodology section below for a fuller description of AR). However, in past research some of the variables used to assess IF in AR were so highly derived that their validity was questionable. In addition, all previous studies were in the USA and this was the first study in the UK, where different norms apply.

In 2003, Paul reported AR reading practice quiz data from 50,823 students in grades 1–12 in 2365 classrooms in 139 schools across 24 states, who read over three million books. Average Percent Correct (APC) on quizzes was a significant predictor of reading test scores at all levels of ability. Students with high engaged time *and* high quality of engagement had high reading achievement gains. However, students with high engagement time but low quality of engagement gained little. Thus, practice did not make perfect; rather, high quality and successful practice made perfect.

Borman and Dowling (2004) conducted a multilevel analysis of student and classroom effects on reading achievement with 45,108 students and 2434 classrooms, analysing elementary schools, middle schools and high schools separately. The amount of text that an elementary or middle-school child read was a key predictor of literacy development at all levels of ability. A high success rate over the school year predicted better outcomes at the end of the year. Students who read books that were on average above their ability level performed better on the reading test than students who read books within their optimum reading range. Classroom-level variability was substantial. The study concluded that time and challenge were the key components of independent reading that contributed to growth in overall reading ability.

A more recent report (Renaissance Learning, 2012) analysed data from 2,284,464 students in all 50 states in the USA who used AR and completed a STAR reading pre- and post-test. These students took 112,763,895 quizzes. Quality was estimated by APC, but time spent reading (Engaged Reading Time − ERT) was estimated in a more convoluted manner. For each quiz taken by a student, student performance was evaluated in relation to the length of the book and the number of items correct on the quiz, which was then used to calculate an estimate of ERT. This variable was clearly highly derived. Turning to challenge, to help match students to appropriate reading materials AR gives information about book difficulty levels using the ATOS formula (see Measures for more details) and student difficulty targets in terms of Zone of Proximal Development (ZPD) (a zone around the student's reading ability).

A multiple regression analysis explored whether these factors accounted for a significant amount of variance in STAR Reading gains across the school year. STAR Reading post-test scores were regressed onto APC, ERT, and ZPD while controlling for pre-test scores. The regression beta coefficients were for APC 3.07 ($t = 319.53$), for ERT 0.54 ($t = 64.21$) and for Percent Quizzes Passed Within or Above ZPD 1.36 ($t = 172.98$) (all statistically significant, given the large sample size, but also large). Thus of the implementation factors, APC appeared the most influential, and ERT the least.

In summary, Paul (2003) found that students with high time spent reading *and* high average percent correct did best on reading tests. Borman and Dowling (2004) found that the amount of text read and a high level of challenge were associated with high reading test results. Renaissance Learning (2012) found that APC was highly related to reading test results, challenge quite highly related, but reading time much less related. Although the derivation of reading time was doubtful, all these studies measured implementation over a full year.

The following study took a somewhat similar approach to the Renaissance Learning (2102) study, but on data from the UK with different norms, contrasting a theoretical approach proposing APC and ERT as the key variables to an empirical approach exploring all the variables embodied in the AR feedback.

4. Research questions

1) Is better performance in key variables in the implementation quality of AR associated with better reading outcomes?
2) Does a theoretical model utilising highly derived variables of IF show a more complete relationship between IF and outcomes, or does an empirical model?
3) Do primary/secondary status, gender and socio-economic status influence these findings?

5. Method

5.1. Sample

The sample comprised all students in the UK for whom AR and STAR results were available for the academic year (n = 852,295 in 3243 schools). This was 10.15% of the 8.4 million children in UK schools (Department for Education, 2015). Schools using AR only in primary numbered 1036 and schools using AR only in secondary numbered 1604. The number of schools using AR in both primary and secondary sectors (including middle schools and special schools) was 603, and in each case each year of students was allocated to primary or secondary as appropriate. Students in high schools outnumbered students in primary schools by three to one. However, data were not available on all variables for all students, since some schools were included which did not provide pre-post test scores for all classes/years of student in the school. Consequently, some analyses were conducted on considerably fewer students. However, the number of students for each analysis was always large and is noted in the text.

5.2. Measures

5.2.1. STAR reading

STAR (Standardized Test for the Assessment of Reading) is a computerised standardized (norm-referenced) adaptive item-banked reading test. Pupils respond to sentences followed by multiple-choice questions on a computer screen. The test is adaptive, i.e. it responds to the performance of each individual student. If the pupil succeeds, harder questions are given. If the pupil fails, easier questions are given. This greatly reduces testing time and student stress. The test is also item-banked, i.e. it has multiple items at the same level. Consequently students cannot copy from each other as no-one is doing the same test. This also enables the test to be taken frequently without practice effects. On completion feedback is available immediately to the teacher and/or pupil. STAR Reading has test-retest reliability of 0.92, split-half reliability of 0.91 and generic reliability of 0.97 in the US. Generic reliability in the UK was 0.94. In term of validity, STAR Reading correlates at 0.96 with the Degrees of Reading Power test. Predictive and concurrent validity with a great number of other reading tests are reported (Renaissance Learning, 2014a, 2013).

5.2.2. Accelerated reader

Accelerated Reader (AR) is a personalized practice and progress-monitoring system that helps teachers accurately and efficiently monitor pupil progress in quantity, difficulty and comprehension (quality) of books read. First, a pupil chooses and reads a book at school, at home, elsewhere or at a combination of these. Each book clearly shows its level of difficulty, based on the ATOS readability formula which assesses the whole book (Milone, 2014) and is known to be a valid and reliable estimate of text complexity (Nelson, Perfetti, Liben, & Liben, 2012). (Begeny and Greene (2014) assessed readability formulae and found them somewhat unreliable, especially for lower ability students, but they did not include ATOS.).

Next, the pupil takes a computerised quiz of 5, 10, or 20 questions depending on the length of the book. Then, pupil and teacher receive immediate computerised feedback with reports detailing books read, number of words read, book reading level and level of comprehension (Percent Correct on the quiz). Each quiz may only be taken once. The questions are sufficiently detailed to ensure that students who have only “seen the film” or heard about the book from a peer cannot pass the test. The reliability of the quizzes is quite low when the quizzes are short, but rises to 0.77 (Cronbach Alpha) for 10-question quizzes and 0.89 for 20-question quizzes (Renaissance Learning, 2014b). Composite reliability for 10 quizzes rose to 0.998. A study of students completing AR quizzes on books they had not read showed that while 10% of them guessed successfully on 5-word quizzes, this fell to 2% on 10-word quizzes and 0.1% on 20-word quizzes.

Essentially AR is a type of formative assessment, since the results of each quiz performance have indications for how the student should approach the next book. AR was designed to make the job of managing book reading easier and more reliable, whilst also motivating pupils to read more books for pleasure. The formative feedback helps teachers shape subsequent reading instruction, guide individual pupils and motivate children to continue reading. Definitions of terms used in this paper are now offered.

5.3. Definitions

5.3.1. Achievement

STAR Scaled Score (SS) ranges from 0 to 1400 and spans years 1–13. It is based on the difficulty of the questions and the number of correct responses.

STAR Student Growth Percentile (SGP) (Betebenner, 2011) is taken from SS scores on two or more tests within 18 months to give an indication of the student's growth trajectory. SGP is a norm-referenced percentile-based index ranging from 1 to 99. It indicates how exemplary a student's growth from one test window to another is relative to students in the same grade with a similar achievement history across the US. SGP indicates past growth trajectory and predicts future growth trajectory. Because SGP is a mathematical manipulation, normal issues of reliability and validity do not apply, but issues of accuracy and precision do. Shang, VanIwaarden, and Betebenner (2015) found that SGP tends to overestimate among students with higher prior achievement and underestimate among those with lower prior achievement, affecting 10% of students. Wright (2010) noted that SGPs correlated highly with value-added models but both under-estimated high-poverty classrooms, with SGP under-estimating least. The simulation-extrapolation method (SIMEX) was used to correct these anomalies.

5.3.2. Implementation

AR Average Percent Correct (APC) is the percent of correctness of the student's answers to the quiz questions, in this case aggregated over all books the student has read.

AR AverageBookLevel-MidGP (ABL-midGP) was a derived variable intended to indicate the degree of challenge in the books each student was reading. The ABL was determined by the ATOS formula, aggregated yearly. From this was subtracted the chronological age (or more precisely, the Grade Placement in years and months) of each student.

5.3.3. Other

Pupil Premium is additional funding for publicly funded schools in England intended to raise the attainment of disadvantaged pupils and close the gap between them and their peers. It is allocated regarding students who have been eligible for free school meals at any point in the last six years, children who are looked after by the local authority, and for children whose parents are currently serving in the armed forces. The Percentage of pupils in the school for whom the premium is received is the variable (PPP).

5.4. Data analysis

Non-parametric analyses were carried out on categorical variables which did not meet the assumptions for parametric tests, particularly Kruskal-Wallis X^2, Mann-Whitney U and Chi-squared χ^2. For the exploratory correlation analysis the Pearson product-moment correlation r was used. For the non-categorical variables, parametric analysis was carried out using Student's t-test for independent samples, Linear Regression Analysis R^2, Forced-entry Multiple Regression Analysis R^2 and subsequent ANOVAs F. Because the sample was not randomly selected the statistical significance of comparisons was not emphasized. In any event the sample was so large that all comparisons except one reached statistical significance (even when the difference was quite small). Variance Inflation Factors (VIFs) were calculated for each regression computation. For the simple linear regressions and the multiple regression computations these were generally less than one, indicating no multicollinearity.

6. Results

In these Results, we will first examine relationships between variables in an exploratory analysis. We will then examine the relationship between theoretical AR implementation variables and achievement. We will then examine the relationship between empirical AR implementation variables and achievement. Further analysis of primary-secondary and gender differences are then reported.

6.1. Exploratory analysis

Pre-, *mid*- and post-scores were very highly correlated for SSGain and SGP (e.g. r = 0.884, 0.885). Consequently the decision was taken to disregard Mid scores. Outcome measures tended to be highly correlated with each other, as did implementation measures, but outcome measures and implementation measures were not highly correlated with each other.

6.1.1. Primary vs. secondary

Five variables were at the ratio scale of measurement. Q-Q plots for these variables confirmed normality and indicated homoscedasticity. Student's t-test was used to analyse the differences between primary and secondary children. The distribution of AR Implementation Category was categorical, peaking at category 2, and was analysed with Mann-Whitney.

Primary children did far better on reading outcomes than secondary school children. (This is the gain on SS within the year, not just the absolute score on SS.) Primary school pupils also did far better on APC than secondary school pupils. They also did much better on ABL-midGP, yielding the largest t in the table. AR Implementation Category was more modestly ahead for

primary pupils. Finally, Pupil Premium was higher at secondary level in England, which means these schools were dealing with a higher level of socio-economic disadvantage than the primary schools, which may partially account for some of the other differences (see Table 1).

6.1.2. Gender

Analysis of gender differences shows marked differences between genders on outcome and implementation measures (Table 2). Five variables were at the ratio scale of measurement. Q-Q plots for these variables confirmed normality and indicated homoscedasticity. Student's t-test was used to analyse the differences between primary and secondary children. The distribution of AR Implementation Category was categorical, peaking at category 2, and was analysed with Mann-Whitney.

Males were significantly worse on outcome measures SSGain and SGP. They were also significantly worse on implementation measures APC, ABL-midGP and ARImplementationCategory. However, in England the genders were the same in terms of their degree of disadvantage as measured by Pupil Premium.

6.1.3. Effect of reading ability in relation to age

The relationship between reading ability in relation to age and indicators of gain and implementation is a topic of interest. We might expect that higher reading ability students (who presumably also have higher reading motivation) would show higher IF than lower reading ability students. We might also expect that higher reading ability students (who have probably read more books within AR and might well be reading more widely outside of AR) might have difficulty showing additional gains in reading, as their performance is already high. Thus lower reading ability students might actually show higher reading gains. PreReadingAbilityAge correlated positively with both IF indicators: APC (0.339, n = 561395) and ABL−midGP (0.591, n = 409792), but not with achievement (SSGain r = −0.106, n = 446491; SGP r = 0.049, n = 439373). Thus students of higher reading ability in relation to age did implement AR at a higher level than students of lower ability. Despite this, they gained in reading at a lower level than students of lower ability.

6.1.4. Socio-economic status (SES) and pre-test scores

SES is often seen as a major determinant of performance. Our measure of SES (in England) was Pupil Premium Percentage (PPP). However, when we looked at the relationship between SSGain/SGP and PPP, we found there was hardly any relationship at all. The correlation between PPP and SSGain was −0.029, n = 394797; between PPP and SGP $r = -0.054^{**}$, n = 387828. Even more surprisingly, there was also little correlation between PPP and pre-test or post-test scores (PPP x SSPre $r = -0.163$, n = 554222; PPP x SSPost $r = -0.168$, n = 519507). The lack of linkage between gain and PPP as an indicator of socio-economic status (SES) is good news, since it implies that the effects of AR are not determined by SES.

6.2. Theoretical investigation of IF

The developers of the AR software recommend that AR be implemented with APC above 85% for each student and with time devoted to silent reading within the school class day. AR Implementation was therefore categorized utilising APC and ERT in the class into:

1. Virtually no AR use
2. Low AR use (below 85% APC or less than 15 min/day ERT)
3. Moderate AR use (85% APC or higher and 15−29 min/day ERT)
4. Best practice AR use (85% APC or higher and 30 + min/day ERT).

The rationale for this is in Renaissance Learning (2012). These categories were related to median SGP scores. The median was used to avoid the effect of outliers. With such a large data set, the difference between the mean and the median is insignificant - it is only with smaller data sets based on an individual school or classroom that differences can arise between the median and the mean.

Table 1
Effect of primary/secondary status on outcome and implementation measures.

	Primary			Secondary			t	U
	n	mean	sd	n	mean	sd		
SSGain	109292	103.05	110.908	308752	70.07	151.337	66.053	
SGP	109163	51.51	29.885	302109	47.04	28.883	43.366	
APC	204602	77.683	15.823	478125	72.731	18.502	105.656	
ABL-MidGP	161557	-0.996	1.207	362085	−3.654	1.262	713.517	
Pupil Premium	176222	23.010	18.395	461273	26.940	16.966	−80.836	
ARImplementationCategory	205756	2.280	0.623	480779	2.160	0.510		78.700

All $p < 0.001$.

Table 2
Effect of gender on outcome and implementation measures.

	Female			Male			*t*	*U*
	n	mean	sd	n	Mean	Sd		
SSGain	187514	80.390	138.837	194356	73.970	147.206	13.870	
SGP	184860	49.160	28.584	191474	46.730	29.810	25.521	
APC	230663	75.926	16.908	241834	73.619	17.769	45.694	
ABL-MidGP	180370	−2.769	1.685	187770	−2.852	1.745	14.702	
Pupil Premium	248235	25.960	17.385	260648	25.910	17.358	0.982	
ARImplementationCategory	262558	2.120	0.715	276166	2.040	0.623		45.919

All $p < 0.001$ except Pupil Premium $p = 0.326$.

Overall we found a clear positive relationship between AR Implementation Categories and SGP (Fig. 1). A Kruskal-Wallis test revealed a difference in SGP across the different AR implementation groups, X^2 (3, 445931) = 4967.097. Follow up Mann-Whitney *U* tests revealed significant differences between all of the groups ($U = 7.656–55.878$ in all comparisons). A χ^2 test comparing the frequencies in each implementation category with normal expectations (a flat distribution) found a very large $\chi^2 = 695844$ (d.f. = 3).

When we looked for differences between Primary and Secondary, we found that the pattern was essentially the same, although Primary did show more elevated results than Secondary (Fig. 2). For primary school, a Kruskal-Wallis Test revealed a difference in SGPs across the different AR implementation groups, X^2 (3, 111650) = 1567.089. Follow up Mann-Whitney *U* tests revealed significant differences between all of the groups, ($U = 5.616–31.750$). For secondary school, a Kruskal-Wallis test revealed a difference in SGPs across the different AR implementation groups, X^2 (3, 334281) = 3363.766. Follow up Mann-Whitney *U* tests revealed significant differences between all of the groups, ($U = 6.264–45.905$ in all comparisons). A χ^2 test comparing the actual frequencies with normal expectations (a flat distribution) found a very large χ^2 for both primary (149853) and secondary (562157) (d.f. = 3 in both cases).

When we looked at the picture for struggling readers (percentile rank below 25), we found the same pattern (Fig. 3). A χ^2 test comparing the actual frequencies with normal expectations (a flat distribution) found a very large χ^2 (278018), d.f. = 3.

When we looked at students who had free or reduced lunch (a measure of low SES), although the numbers were lower because this data was not always available, we found the same (Fig. 4). A χ^2 test comparing the actual frequencies with normal expectations (a flat distribution) found a very large χ^2 (32452, d.f. = 3).

When we looked at the pattern for students who were learning English as a second or additional language, we found the same (Fig. 5). A χ^2 test comparing the actual frequencies with normal expectations (a flat distribution) found a very large χ^2 (13499, d.f. = 3).

Of course, this analysis is based on a theory about how AR works, but it has the problem that it divides the data up into categories (and then adds two categories together). This is wasteful of data. An analysis which uses all the data in continuous variables is likely to be much more analytical in uncovering relationships. Additionally, all variables can be analysed, rather than just the ones theory predicts to be important.

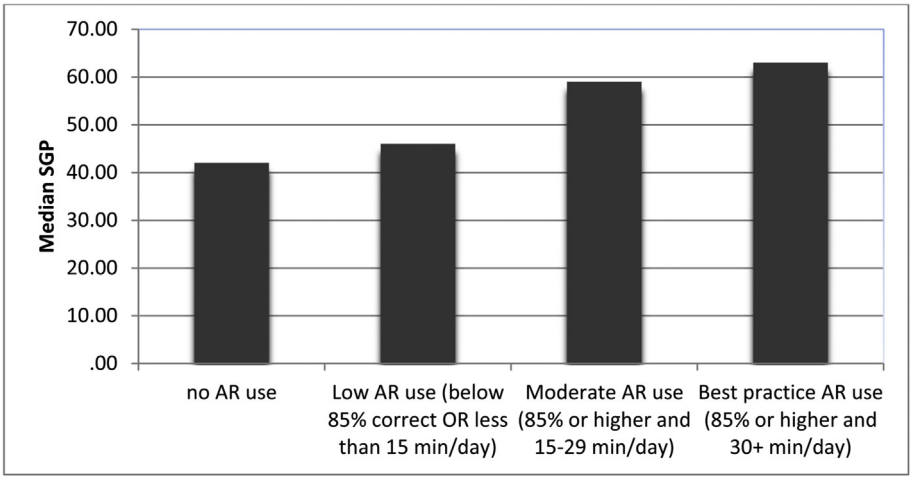

Fig. 1. AR implementation categories and SGP overall.

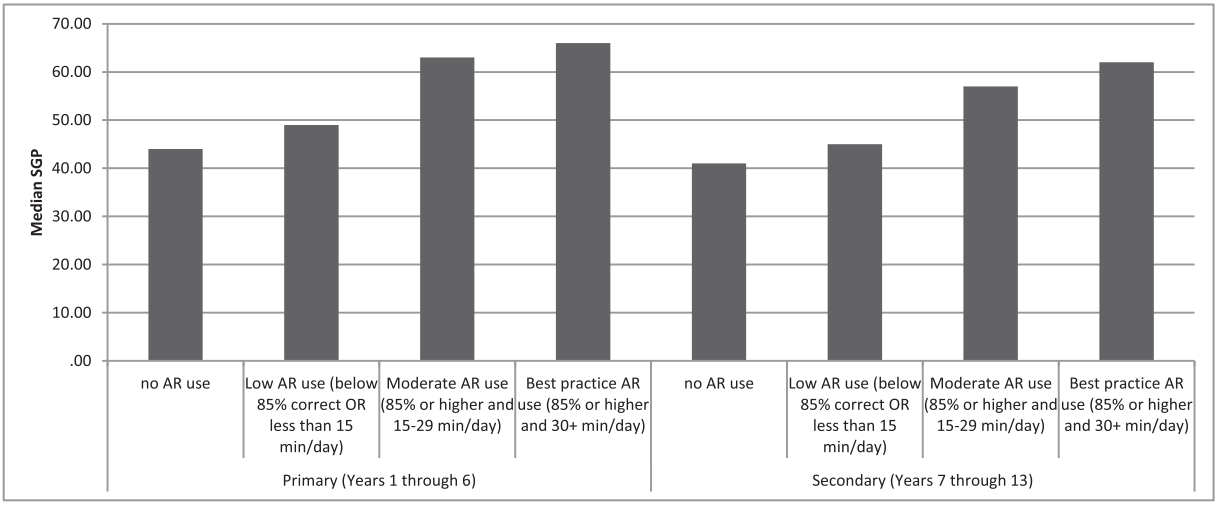

Fig. 2. AR implementation categories and SGP for primary and secondary.

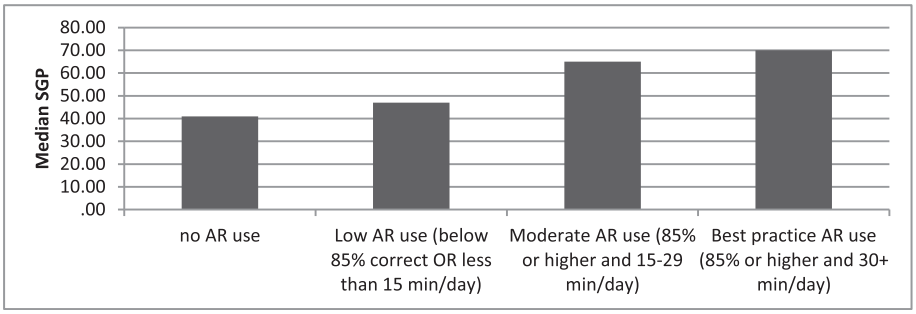

Fig. 3. AR implementation categories and SGP for struggling readers.

6.3. Empirical investigation of IF

6.3.1. Linear regression on implementation and outcome variables

SSGain and SGP were normally distributed and homoscedastic. The implementation variables of interest included APC, ABL-MidGP and ARImplementationCategory, but also variables such as Quizzes Taken, Quizzes Passed, TotalPointsEarned, AverageBookLevel and EngagedReadingTime. APC and ABL-MidGP were both were normally distributed and homoscedastic.

We calculated a simple linear regression for each implementation variable independently. The results are in Table 3. We noted that the numbers of students for whom these data was available was considerably reduced from the original sample, but a sample of almost 350,000 could be considered adequate. Variance accounted for in SSGain was: APC = 2.2%, ABL-MidGP = 1.8%, total 4.0%. In SGP variance accounted for was: APC = 4.1%, ABL-MidGP = 2.4%, total 6.5%. These proportions were small but for other variables the proportions were much smaller. It seems that there was much random variation in the data, i.e. they were noisy.

6.3.2. Multiple regression on implementation and outcome variables

Calculating a forced-entry multiple regression for SGP, we found that APC and ABL-midGP accounted for the majority of the variance - 0.041 (4.1%) and 0.014 (1.4%) respectively, making a total of 5.5%. ANOVAs were $F(1, 348419) = 14858.250$ and $F(1, 348418) = 10169.483$, significance was high (APC $t = 106.575$, ABL $t = 72.502$), and VIF was 1.038. Other variables all accounted for less than 0.5. For SSGain it was a little lower (0.022 and 0.011 respectively, accounting for 2.2% and 1.1% of the variance, totalling 3.3%). Significance was again high (ANOVA $F(1, 354489) = 7859.695$ and $F(1, 354488) = 6088.841$) and VIF was 1.037. Other variables accounted for 0.003 (0.3% of the variance). The numbers of students for whom data were available was considerably reduced from the original sample, but almost 350,000 could be considered adequate for this purpose.

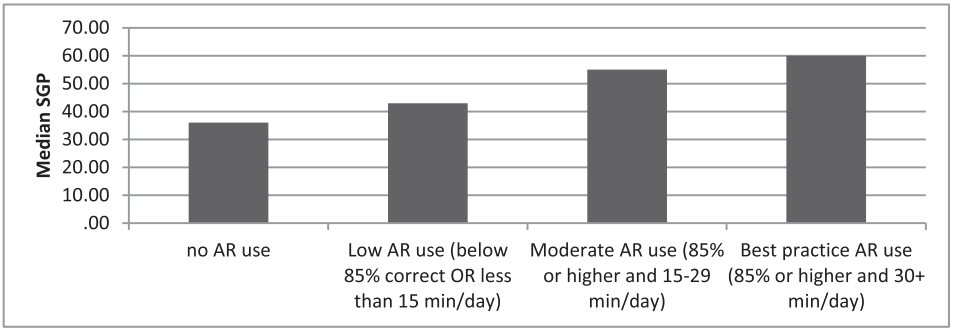

Fig. 4. AR implementation categories and SGP for free/reduced lunch students.

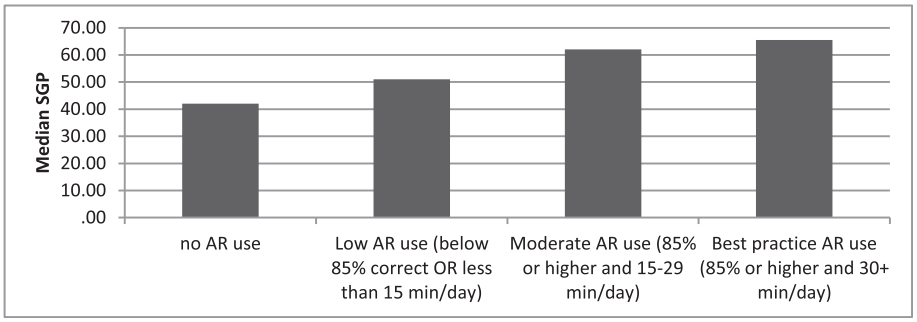

Fig. 5. AR implementation categories and SGP for students learning English as a second or additional language.

Table 3
Linear Regression of Gain x Implementation Variables.

	SS Gain	SGP
AvgPercentCorrect	$R^2 = 0.022 = 2.2\%$ of Variance explained. $F(df = 1,416477) = 9516.722$ t(ARImplementationCategory) = 97.554	$R^2 = 0.041 = 4.1\%$ of Variance explained. $F(df = 1,409790) = 17453.558$ t(AvgPercentCorrect) = 132.112
AvgBookLevel-MidGP	$R^2 = 0.018 = 1.8\%$ of Variance explained. $F(df = 1,354489) = 6417.207$ t(AvgBookLevel-MidGP) = 80.107	$R^2 = 0.024 = 2.4\%$ of Variance explained. $F(df = 1,409790) = 8697.182$ t(AvgBookLevel-MidGP) = 93.259
AR Implementation Category	$R^2 = 0.012 = 1.2\%$ of Variance explained $F(df = 1,451456) = 5564.736$ t(ARImplementationCategory) = 74.597	$R^2 = 0.020 = 2\%$ of Variance explained. $F(df = 1,444269) = 8973.131$ t(ARImplementationCategory) = 94.727
QuizzesTaken	$R^2 = 0.007 = 0.7\%$ of Variance explained. $F(df = 1,416477) = 3098.097$ t(QuizzesTaken) = 55.661	$R^2 = 0.005 = 0.5\%$ of Variance explained. $F(df = 1,444269) = 1865.875$ t(QuizzesTaken) = 43.196
QuizzesPassed	$R^2 = 0.013 = 1.3\%$ of Variance explained. $F(df = 1,416477) = 5438.868$ t(QuizzesPassed) = 73.749	$R^2 = 0.012 = 1.2\%$ of Variance explained. $F(df = 1,409790) = 4910.814$ t(QuizzesPassed) = 70.078
TotalPointsEarned	$R^2 = 0.008 = 0.8\%$ of Variance explained. $F(df = 1, 416477) = 3369.776$ t(TotalPointsEarned) = 58.850	$R^2 = 0.021 = 2.1\%$ of Variance explained. $F(df = 1,409790) = 8886.220$ t(TotalPointsEarned) = 94.267
AvgBookLevel	$R^2 = 0.000 = 0\%$ of Variance explained. $F(df = 1,405151) = 161.464$ t(AvgBookLevel) = 12.707	$R^2 = 0.011 = 1.1\%$ of Variance explained. $F(df = 1,398634) = 4332.853$ t(AvgBookLevel) = 65.824
EngagedReadingTime	$R^2 = 0.010 = 1.0\%$ of Variance explained. $F(df = 1,416477) = 4201.135$ t(fERT_RP) = 64.816	$R^2 = 0.018 = 1.8\%$ of Variance explained. $F(df = 1,409790) = 7680.559$ t(fERT_RP) = 87.639

All $p < 0.001$.

6.3.3. Multiple regression on primary/secondary and gender

Primary/Secondary and Gender status made substantial differences to outcomes and implementation in the exploratory analysis. Consequently we undertook multiple regressions on these variables separately. Results are in Table 4. The numbers

Table 4
Multiple Regression of Gain x Primary/Secondary and Gender Variables.

	SS Gain	SGP
Primary	$R^2 = 0.047 = 4.7\%$ of Variance explained	$R^2 = 0.077 = 7.7\%$ of Variance explained
	F (APC)(df = 1,99073) = 3215.001	F (APC)(df = 1,98941) = 5675.471
	F(ABL)(df = 1,99072) = 2443.328	F(ABL)(df = 1, 98940) = 4105.477
	t(constant) = 7.328	t(constant) = 34.174
	t(APC) = 48.807	t(APC) = 65.984
	t(ABL) = 40.239	t(ABL) = 48.967
	VIF = 1.035	VIF = 1.035
	SSGain = 15.728 + 1.258(APC) + 12.026(ABL)	SGP = 19.484 + .452(APC) + 3.887(ABL)
Secondary	$R^2 = 0.019 = 1.9\%$ of Variance explained	$R^2 = 0.047 = 4.7\%$ of Variance explained
	F (APC)(df = 1,255414) = 4138.562	F (APC)(df = 1,249476) = 8663.009
	F(ABL)(df = 1,255413) = 2415.843	F(ABL)(df = 1,249475) = 6093.124
	t(constant) = 5.258	t(constant) = 99.998
	t(APC) = 60.474	t(APC) = 85.164
	t(ABL) = 26.117	t(ABL) = 58.353
	VIF = 1.017	VIF = 1.018
	SSGain = 9.549 + 1.143(APC) + 6.531(ABL)	SGP = 34.277 + .306(APC) + 2.785(ABL)
Male	$R^2 = 0.033 = 3.3\%$ of Variance explained	$R^2 = 0.054 = 5.4\%$ of Variance explained
	F (APC)(df = 1,152274) = 3317.586	F (APC)(df = 1,149828) = 5963.196
	F(ABL)(df = 1,152273) = 2586.858	F(ABL)(df = 1,145620) = 4286.757
	t(constant) = 7.663	t(constant) = 68.633
	t(APC) = 49.328	t(APC) = 67.422
	t(ABL) = 42.621	t(ABL) = 50.104
	VIF = 1.033	VIF = 1.034
	SSGain = 15.699 + 1.182(APC) + 9.204(ABL)	SGP = 28.811 + .331(APC) + 2.216(ABL)
Female	$R^2 = 0.036 = 3.6\%$ of Variance explained	$R^2 = 0.058 = 5.8\%$ of Variance explained
	F (APC)(df = 1,147925) = 3457.862	F (APC)(df = 1,145621) = 6816.899
	F(ABL)(df = 1,147924) = 2736.462	F(ABL)(df = 1,145620) = 4464.638
	t(constant) = 9.143	t(constant) = 62.945
	t(APC) = 49.228	t(APC) = 72.368
	t(ABL) = 44.374	t(ABL) = 44.924
	VIF = 1.041	VIF = 1.042
	SSGain = 19.141 + 1.181(APC) + 9.511(ABL)	SGP = 27.382 + .361(APC) + 2.002(ABL)

All $p < 0.001$.

of students for whom data was available was considerably reduced from the original sample, but 350,000 could be considered adequate for this purpose.

For SGP, 7.7% of the gain was accounted for in Primary, and 4.7% in Secondary. SSGain accounted for 4.7% of the variance for Primary, but only 1.9% for Secondary. Thus Primary showed higher accountability for variance than Secondary. For Gender SGP, 3.4% of the gain was accounted for by Males, and 5.8% for Females. For SSGain, 3.3% was accounted for by Males and 3.6% for Females. Females accounted for more of the variance than males.

6.4. Summary

AR IF is positively related to STAR achievement gain (as both SSGain and SGP), whether defined theoretically (APC and ERT) or empirically (APC and ABL-midGP), although the empirical association is stronger. In both the exploratory and regression analyses, Primary do far better than Secondary children on both implementation quality and reading outcomes (although secondary schools are dealing with a somewhat higher level of socio-economic disadvantage in England). Similarly, girls do far better than boys in both analyses. Students of higher reading ability in relation to age implemented AR at a higher level than students of lower ability, but gained in reading at a slightly lower level. There was hardly any association between reading outcome gains and Pupil Premium, and also little correlation between Pupil Premium and pre-test or post-test achievement - the effects of AR combination were not determined by SES.

7. Discussion

7.1. Connection to previous literature

Most previous literature focused on direct and indirect measures of IF outside of book reading, so here we are mostly trying to connect computer-based data on book reading with a somewhat disparate set of measures.

Previous research on the relationship between IF studied indirectly (by asking teachers for their subjective opinion) and student outcome was not encouraging - generally indirect measures related poorly to achievement tests. However, Noltemeyer et al. (2014) found a positive relationship with outcomes, but this was in the context of Response to Intervention initiatives, which perhaps are more highly structured than other forms of teaching. On the other hand, Balu et al. (2015) did

not find such a relationship in the context of RTI interventions. Sharp et al. (2016) also found a positive relationship between IF and outcomes in elementary schools. Mostly these studies focused on teacher behaviour.

Regarding direct observational measures, there was considerable variability across teachers, and generally IF likewise did not impact on outcomes. However, these studies offered more clues as to possible mediating variables. McIntyre et al. (2005) found high implementers had much support; a practical, clear model; extensive professional development; or a combination of these. Henninger (2010) noted low complexity led to high implementation, high acceptability led to high implementation. Both led to high outcomes. Arguably the use of AR and STAR lowers complexity, in that the process of testing, scoring and feedback is automated. However, there is still considerable complexity for the teacher in interpreting the feedback and taking action based upon it.

Mcintyre et al.'s (2005) "practical model" and Henninger's (2010) "low complexity" seem to have a good deal in common. Benner, Nelson, Stage, and Ralston (2011) found two teacher behaviours (following the lesson format as designed and re-teaching lessons when needed) were particularly related to outcomes. This study has implications for future research, in terms of finding the most effective teacher behaviours over many contexts. However, again many of these studies focused on teacher behaviour. Fogarty et al. (2014) emphasized the importance of measuring multiple dimensions of IF – and their study did include student responsiveness and found fidelity was related to outcomes. Additionally, most direct and indirect studies focused on a relatively short period of implementation.

Neither direct nor indirect methods included any computer-based student-response measures of IF or outcome. Turning to previous reports on AR/STAR, we find that APC was noted as the strongest variable (e.g. Borman & Dowling, 2004; Paul, 2003), which result we have replicated here. Concerning time and challenge, which previous research found related to outcomes, the variables used were highly derived and of uncertain reliability (e.g. Renaissance Learning, 2012). The present research did not replicate the finding regarding ERT. Nor did it replicate the finding regarding challenge using the methods from the previous literature. It did however replicate this finding using a simpler metric more closely tied to the AR feedback.

7.2. Interpretation

These results suggest a causative reason for boys doing worse than girls on outcomes – boys are not implementing as well as girls. Thus they do not read as much text (either fiction or non-fiction), they do not pay enough attention to the comprehension of the text, and they tend to choose books which are too easy for them. This suggests an overall lack of motivation for reading books. It may have to do with the ethos surrounding reading in secondary; boys in particular not seeing it as "cool". This applies to non-fiction as well as fiction (Topping, 2015).

Likewise, the dissimilar results from primary and secondary schools are striking. After secondary transfer, both boys and girls read less text, comprehend it less well, and choose books that are too easy for them as they become older. Researchers may ascribe this to organizational changes on passing from primary to secondary, while teachers may ascribe it to maturational changes as children grow older. Perhaps children feel they have learned to read in primary school and need not enhance their reading ability.

Direct and indirect studies emphasising teacher behaviour have shown weak association with outcomes, while the present student-response study also shows a small amount of variance accounted for. A future study which collected data from indirect methods, direct methods and computerised methods simultaneously on the same students would allow the comparison of all three methods with outcomes on reading tests. Then some combination of these methods which had the greatest predictive effect could be established.

Further, this comparison of methods with outcomes could be extended to subjects other than book reading. Mathematics is the most obvious example – a subject in which Renaissance Learning offers a computerised quiz system and norm-referenced mathematics test, just as with reading.

Considering the implications from the previous literature, a number of characteristics are likely to be associated with reliability of measurement and high IF:

- higher structure in the intervention
- simplicity and clarity in the model of intervention
- greater support with continuing professional development and onward coaching of teachers
- low complexity and high acceptability by the teachers
- specification of the most predictive components
- investigating student as well as teacher behaviour
- measurement over a period of at least a year.

Finally, the variable found by Benner et al. (2011) needs emphasising – re-teaching of material inadequately learned. Previous studies tended to assume the intervention proceeded in a linear manner as designed, irrespective of student response. However, student response is critical and many teachers evaluate and re-evaluate their students' response and re-teach (often in a different way) when necessary. Benner et al. (2011) was the only study to mention this.

7.3. Limitations

The present study had a number of limitations, as well as a number of advantages. The principal advantage was the large sample size. This led to de-emphasis on statistical significance, since almost everything was statistically significant. The decision to discard the midpoint test results markedly reduced the sample size. It was also further reduced by some schools entering all pupils for testing but then not providing results on them all. Nonetheless, the sample size remained large, certainly much larger than in most studies of reading.

SGP tends to under-estimate schools in socio-economically disadvantaged areas and over-estimate schools in advantaged areas. This suggests that when interpreting the tables, readers should judge flexibly in the top and bottom quintiles.

Outcome measures were highly correlated with each other at pre and post. Implementation variables were also highly correlated with each other. But outcome variables and implementation variables were not highly correlated with each other. Regression coefficients were small, implying the variables investigated accounted for a small percentage of the variance, but this was even truer of other variables. It seems there was a large amount of noise in the data.

7.4. Implications for practice, policy and future research

7.4.1. Practice

Teachers should strive to maximize implementation of what appear to be the major determinants of higher outcomes from the empirical analysis – APC and Challenge (ABL—midGP). Of course, teachers are working indirectly with individual students who generate the data, so much of their work will involve explaining to students and subsequently coaching them. At a systemic level, when teachers evaluate the success of AR in their schools, they should carefully consider the evidence on these two key indicators of IF as well as the level of student outcomes, and strive to increase them.

Beyond this narrow focus on computer assessment of book reading, teachers and related practitioners should take extreme care when evaluating the results of randomised controlled trials, to ensure that adequate evidence of implementation fidelity is also available, and that it is of adequate reliability.

7.4.2. Policy

Policy-makers (including school inspectors) at local and national level should carefully consider the evidence on these two key indicators of IF as well as the level of student outcomes. Policy-makers need to be sharply aware that randomised controlled trials without accompanying reliable evidence of implementation integrity are of little value, and should not be over-interpreted. The advice that they give to teachers should reflect this caution. They may consider providing relevant professional development opportunities to teachers and schools. It will be important that any findings are made available to the wider public.

7.4.3. Future research

Should studies similar to this be repeated, it would be useful to investigate the two key empirically determined variables (APC and ABL-midGP). In the UK, further work could seek to incorporate the mid-year STAR data. As noted above, a further study of indirect, direct and computerised methods of establishing IF with the same pupils would be highly valuable in reading and could be extended to other subjects, particularly mathematics. Considering the wider field, future researchers may wish to establish studies which investigate the effects of some of the other variables mentioned. An example would be the investigation of mediating variables. Overall, researchers should never simply focus on the outcomes of randomised controlled trials without satisfying themselves that reliable evidence of IF is also provided.

8. Conclusion

AR implementation quality (APC and ABL-midGP) is positively related to STAR achievement gain (SSGain/SGP). Empirical analysis showed some stronger relationships than theoretical analysis. Primary schools did better on IF and outcomes than secondary schools. Males did significantly worse on IF and outcomes than females, although there was no difference between genders in socio-economic disadvantage. Students of higher reading ability in relation to age implemented AR at a higher level than students of lower ability, but gained in reading at a slightly lower level. The effect of AR implementation as reflected in STAR results was not affected by socio-economic status.

Thus in relation to the research questions, we did find (RQ1) that better performance in key variables in the implementation quality of AR was associated with better reading outcomes. We also found (RQ2) that an empirical model of the relationship between IF and outcomes showed a more complete influence on outcomes than a theoretical model utilising highly derived variables of IF. Primary schools did better than secondary schools on implementation and achievement (RQ3). Females did better than males on implementation and achievement. Socio-economic status was not related to either implementation or achievement.

Indirect, direct and computerised student-response measures of IF all had some problems in predicting pupil outcome. It is suggested that future research needs to triangulate indirect, direct and computerised student-response measures with the

same students over a period of at least a year, to establish which combination might be the most predictive in the longer run. Then the same approach should be attempted with other subjects, probably starting with mathematics, where computerised student process and achievement measures generating large amounts of data are already available.

Clearly, establishing IF in book reading is much harder than might have been expected, if the criterion of validity is relationship to achievement outcomes. Beyond book reading, it seems likely that direct, indirect and computer-based methods will all have limited reliability. Consequently a blended method which incorporates all three, albeit with very different sample sizes given the relative cost of collecting direct and indirect data, may well offer a way forward. Overall, however, collecting reliable IF data will be costly.

Further, beyond book reading, it is not clear whether we can expect measuring IF to be easier or harder. Nor indeed whether higher IF will necessarily lead to higher achievement. Computerised student response measures are not yet available in many other areas of the curriculum, and seem likely to be of even less reliability even if they were. Computerised methods of assessing teacher behaviour seem to be some way in the future. The implications here for the widespread use of randomised controlled trials are considerable. Without reliable IF data, many of them are next to useless, and do not justify the money spent on them. A much larger portion of research resource needs to be devoted to establishing satisfactory multi-component IF measures. Organizations collecting studies on evidence-based interventions need to give much closer attention to the issue of IF.

Conflict of interest

Dr. Topping is a Professor at the University of Dundee. The results of this study do not create a conflict of interest for him.

Financial disclosure

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

References

Balu, R., Zhu, P., Doolittle, F., Schiller, E., Jenkins, J., & Gersten, R. (2015). *Evaluation of response to intervention practices for elementary school reading*. Washington, DC: National Center for Education Evaluation and Regional Assistance. ERIC Number: ED560820. Retrieved from http://files.eric.ed.gov/fulltext/ED560820.pdf. (Accessed 15 June 2016).

Begeny, J. C., Easton, J. E., Upright, J. J., Tunstall, K. R., & Ehrenbock, C. A. (2014). The reliability and user-feasibility of materials and procedures for monitoring the implementation integrity of a reading intervention. *Psychology in the Schools, 51*(5), 517–533.

Begeny, J. C., & Greene, D. J. (2014). Can readability formulas be used to successfully gauge difficulty of reading materials? *Psychology in the Schools, 51*, 198–215.

Begeny, J., Upright, J., Easton, J., Ehrenbock, C., & Tunstall, K. (2013). Validity estimates and functionality of materials and procedures used to monitor the implementation integrity of a reading intervention. *Journal of Applied School Psychology, 29*(3), 284–304.

Benner, G. J., Nelson, J. R., Stage, S. A., & Ralston, N. C. (2011). The influence of fidelity of implementation on the reading outcomes of middle school students experiencing reading difficulties. *Remedial and Special Education, 32*(1), 79–88.

Betebenner, D. W. (2011). *A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth projections/trajectories*. Dover, New Hampshire: The National Center for the Improvement of Educational Assessment. Retrieved from http://www.nj.gov/education/njsmart/performance/SGP_Technical_Overview.pdf. (Accessed 14 June 2016).

Borman, G. D., & Dowling, N. M. (2004). *Testing the reading renaissance program theory: A multilevel analysis of student and classroom effects on reading achievement*. Unpublished manuscript. University of Wisconsin–Madison.

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science, 72*, 40. https://doi.org/10.1186/1748-5908-2-40. Retrieved from https://implementationscience.biomedcentral.com/articles/10.1186/1748-5908-2-40. (Accessed 14 June 2016).

Ciullo, S., Lembke, E. S., Carlisle, A., Thomas, C. N., Goodwin, M., & Judd, L. (2016). Implementation of evidence-based literacy practices in middle school response to intervention: An observation study. *Learning Disability Quarterly, 39*(1), 44–57.

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review, 18*(1), 23.

Darrow, C. L. (2010). *Measuring fidelity in preschool interventions: A microanalysis of fidelity instruments used in curriculum interventions*. Evanston, IL: Society for Research on Educational Effectiveness. ERIC Number: ED514643. Retrieved from http://files.eric.ed.gov/fulltext/ED514643.pdf. (Accessed 15 June 2016).

Department for Education. (2015). *Schools, pupils and their characteristics: January 2015*. London: DfE.

Durlak, J. P., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*, 327–350.

Fedor, L. C. (2013). *The relationship between the level of implementation of scientifically based reading instructional practices in K-3 and grade 3 Pennsylvania system of school assessment reading achievement in north-eastern Pennsylvania*. Ed.D. dissertation. Wilkes-Barre, PA: Wilkes University. ERIC number: ED563676. Retrieved from http://search.proquest.com/docview/1468677472. (Accessed 15 June 2016).

Feldman, J., Feighan, K., Kirtcheva, E., & Heereen, E. (2012). Aiming High: Exploring the influence of implementation fidelity and cognitive demand levels on struggling readers' literacy outcomes. *Journal of Classroom Interaction, 47*(1), 4–13.

Fogarty, M., Oslund, E., Simmons, D., Davis, J., Simmons, L., Anderson, L., et al. (2014). Examining the effectiveness of a multicomponent reading comprehension intervention in middle schools: A focus on treatment fidelity. *Educational Psychology Review, 26*(3), 425–449.

Henninger, K. L. (2010). *Exploring the relationship between factors of implementation, treatment integrity and reading fluency*. Ph.D. Dissertation. Ann Arbor, MA: University of Massachusetts Amherst. ERIC Number: ED519216. Retrieved from http://search.proquest.com/docview/638627099. (Accessed 15 June 2016).

McIntyre, E., Powell, R., Coots, K. B., Jones, D., Powers, S., Deeters, F., et al. (2005). Reading instruction in the NCLB era: Teachers' implementation fidelity of early reading models. *Journal of Educational Research & Policy Studies, 5*(2), 66–102.

Milone, M. (2014). *Development of the ATOS™ readability formula*. Wisconsin Rapids, WI: Renaissance Learning.

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2012). *Measures of text difficulty: Testing their predictive value for grade levels and student performance*. New York: Student Achievement Partners.

Noltemeyer, A. L., Boone, W. J., & Sansosti, F. J. (2014). Assessing school-level RTI implementation for reading: Development and piloting of the RTIS-R. *Assessment for Effective Intervention, 40*(1), 40–52.

Paul, T. D. (2003). *Guided independent reading: An examination of the Reading Practice Database and the scientific research supporting guided independent reading as implemented in Reading Renaissance*. Wisconsin Rapids, WI: Renaissance Learning, Inc.

Renaissance Learning. (2012). *Guided independent reading*. Wisconsin Rapids, WI: Renaissance Learning. Retrieved from http://doc.renlearn.com/KMNet/R005577721AC3667.pdf. (Accessed 16 June 2016).

Renaissance Learning. (2013). *STAR reading technical manual (UK)*. London: Renaissance Learning.

Renaissance Learning. (2014a). *The research foundation for STAR assessments*. Wisconsin Rapids, WI: Renaissance Learning. Retrieved from http://doc.renlearn.com/KMNet/R001480701GCFBB9.pdf . (Accessed 17 June 2016).

Renaissance Learning. (2014b). *Accelerated Reader 360™ understanding reliability and validity*. Wisconsin Rapids, WI: Renaissance Learning. http://doc.renlearn.com/KMNet/R003580612GF885B.pdf. (Accessed 9 March 2017).

Sanetti, L. M. H., & Fallon, L. M. (2011). Treatment integrity assessment: How estimates of adherence, quality, and exposure influence interpretation of implementation. *Journal of Educational and Psychological Consultation, 21*, 209–232.

Schulte, A. C., Easton, J. E., & Parker, J. (2009). Advances in treatment integrity research: Multidisciplinary perspectives on the conceptualization, measurement, and enhancement of treatment integrity. *School Psychology Review, 38*, 460–475.

Shang, Y., VanIwaarden, A., & Betebenner, D. W. (2015). Covariate measurement error correction for Student Growth Percentiles using the SIMEX method. *Educational Measurement: Issues and Practice, 34*(1), 4–14.

Sharp, K., Sanders, K., Noltemeyer, A., Hoffman, J. L., & Boone, W. J. (2016). The relationship between RTI implementation and reading achievement: A school-level analysis. *Preventing School Failure, 60*(2), 152–160.

Stockard, J. (2010). An analysis of the fidelity implementation policies of the what Works Clearinghouse. *Current Issues in Education, 13*(4). Retrieved from http://cie.asu.edu/ojs/index.php/cieatasu/article/view/398/94 . (Accessed 15 June 2016).

Topping, K. J. (2015). Fiction and non-fiction reading and comprehension in preferred books. *Reading Psychology, 36*(4), 350–387. https://doi.org/10.1080/02702711.2013.865692.

Wehby, J. H., Maggin, D. M., Johnson, L., & Symons, F. J. (2010). *Improving intervention implementation and fidelity in evidence-based practice: Integrating teacher preference into intervention selection*. Evanston, IL: Society for Research on Educational Effectiveness. ERIC number ED512822. Retrieved from http://files.eric.ed.gov/fulltext/ED512822.pdf . (Accessed 15 June 2016).

Williams, D. M. (2015). Middle level best practice and student achievement in Texas. *Current Issues in Middle Level Education, 20*(1), 8–17.

Wright, S. P. (2010). *An investigation of two nonparametric regression models for value-added assessment in education*. Cary, NC: SAS Institute Inc. Retrieved from https://education.ohio.gov/getattachment/Topics/Data/Report-Card-Resources/Ohio-Report-Cards/Value-Added-Technical-Reports-1/An-Investigation-of-Two-Nonparametric-Regression-Models-for-Value-Added-Assessment-in-Education-S-Paul-Wright-1.pdf.aspx. (Accessed 16 June 2016).