Implementation Fidelity and Attainment in Computerized Practice of Mathematics

Topping, Keith

Published in:
Research Papers in Education

DOI:
10.1080/02671522.2019.1601759

Publication date:
2020

Document Version
Peer reviewed version

Implementation Fidelity and Attainment in Computerized Practice of Mathematics

Keith J Topping

University of Dundee, Scotland

Measuring the implementation fidelity (IF) or integrity of interventions is crucial, otherwise a positive or negative outcome cannot be interpreted. Direct and indirect methods of IF measurement tend to over-emphasize teacher behaviour. This paper focuses on IF measured by student behaviour collected through computers. Attainment was measured by the STAR test of maths (a computerized item-banked adaptive norm-referenced test). Implementation quality (IF) was measured by Accelerated Maths (AM) (an instruction-free personalized practice and progress-monitoring system in mastery of mathematics skills). Attainment data was gathered in the UK on 20,103 students in 148 schools, and of these implementation data on n=6,285. Only a small percentage of pupils scored on five AM implementation indices at or above the levels recommended. Correlations between attainment and implementation indices were modest, but high implementation was positively correlated with high attainment. Socio-economic status did not appear to affect implementation or attainment. Implementation quality of AM is clearly a problem in the UK, and needs improvement. However, overall students still scored above average on attainment.

Length: 7399 words 28 pages excluding tables (449 words, 4 pages)
Running Head: IMPLEMENTATION FIDELITY IN COMPUTERIZED MATHS

Corresponding author: Professor Keith Topping, School of Education, University of Dundee, Dundee DD1 4HN, United Kingdom. Tel: +44 (0)7854 833556, Email: { HYPERLINK "mailto:k.j.topping@dundee.ac.uk" }

Biographical note: Keith Topping is the Professor of Educational and Social Research at the University of Dundee, Scotland. His research interests are peer assisted learning, parents as educators, computer assisted assessment and early language. He has almost 400 research publications and is translated into 12 languages.

Financial Disclosure: The authors have no relevant financial relationships to disclose.

Data Availability: The data can be made available to other researchers who wish to further analyse them. Contact the author.

Conflict of Interest: The author has no conflicts of interest relevant to this article to disclose.

Implementation Fidelity and Attainment in Computerized Practice of Mathematics

Abstract

Measuring the implementation fidelity (IF) or integrity of interventions is crucial, otherwise a positive or negative outcome cannot be interpreted. Direct and indirect methods of IF measurement tend to over-emphasize teacher behaviour. This paper focuses on IF measured by student behaviour collected through computers. Attainment was measured by the STAR test of maths (a computerized item-banked adaptive norm-referenced test). Implementation quality (IF) was measured by Accelerated Maths (AM) (an instruction-free personalized practice and progress-monitoring system in mastery of mathematics skills). Attainment data was gathered in the UK on 20,103 students in 148 schools, and of these implementation data on n=6,285. Only a small percentage of pupils scored on five AM implementation indices at or above the levels recommended. Correlations between attainment and implementation indices were modest, but high implementation was positively correlated with high attainment. Socio-economic status did not appear to affect implementation or attainment. Implementation quality of AM is clearly a problem in the UK, and needs improvement. However, overall students still scored above average on attainment.

Keywords: mathematics, implementation integrity, implementation fidelity, attainment, implementation, computerized assessment, students, improving classroom teaching

Introduction

Since the emphasis has moved towards "evidence-based" interventions, measuring the quality of implementation has become an increasing preoccupation. There is no point attempting to implement an intervention and measure the outcomes if there is no attempt to see if the method has actually been implemented. Unless IF is assessed, in the case of poor outcome we will not know whether the program did not work or merely was not implemented properly, or both. Even in the case of good outcome, we cannot know whether the program worked and was responsible for the positive outcome.

Implementation fidelity (or integrity) (IF) 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. Schulte, Easton, and Parker (2009) proposed five main elements: adherence to an intervention, exposure or dose, quality of delivery, program differentiation (the extent to which key factors in effectiveness were identified) and participant responsiveness. Schulte et al. (2009) included 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 arose of which of these indices were most related to outcome.

Measuring IF is not easy - researchers find that it is both complex and expensive. Indirect attempts which simply ask teachers if they have implemented well often do not correlate with outcomes. More direct attempts using observational methods (to avoid teacher subjectivity) are expensive (and consequently only usable on a small scale). They also suffer from observer effects – what the teacher did when observed might not have been typical of what they did when not observed. 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.

The Current Paper

Computerized assessment in mathematics generates a large amount of data, which is also gathered over the course of a whole school year. This paper focuses on the effectiveness of differentiated practice in mathematics via the computer-based Accelerated Maths (AM) program and emphasizes student response rather than teacher behaviour. The paper compares and contrasts five different implementation indicators of IF with growth in mathematics attainment on the norm-referenced STAR Math test. The study deploys measures of student response to counter-balance the existing over-emphasis on teacher behaviour. Both attainment and IF measures were 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. The present paper is a companion to previous papers focused on reading (Topping, 2017, 2018).

2.0 Previous Research on IF in Mathematics

This literature review interrogates previous empirical research on IF in mathematics. An initial section explores research on IF in mathematics for programs other than AM. Some of these describe research on indirect measures (self-reports completed subjectively by teachers and head teachers), while others focus on direct measures (completed by observation, although still far from "objective" given possible observer effects). After this, a further section briefly explores the literature on using systems such as STAR Math and AM separately. Then a final section explores in more detail such measures used in conjunction.

2.1 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 “mathematics” AND “implementation fidelity” OR “implementation integrity” OR treatment fidelity” OR “treatment integrity”. The inclusion criteria specified relevance to the research questions and containing empirical data. Fifty-two hits resulted from the search of titles and abstracts. 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 31.

2.2 Implementation Fidelity in Mathematics for Programs Other Than AM

A number of papers on IF in mathematics have disappointing findings - there was no evidence that the intervention was implemented as desired and no evidence of any improvement in attainment. An example is the evaluation of Classroom Assessment for Student Learning (CASL), a widely used professional development program (Randel, Apthorp, Beesley, Clark & Wang, 2016). Schools were randomly allocated CASL or regular professional development. Analysis of 67 schools and 9,596 students yielded no statistically significant impacts of CASL on student mathematics achievement as measured by the state-wide test. No statistically significant impacts were found on teachers' assessment practice.

Why might this be? Holstein (2012) examined teachers' implementation of a mathematical decision-making curriculum. Observations and teacher logs were coupled with interviews and surveys. Four out of six teachers were reasonably faithful to the program. Four types of implementers were identified: (a) "thorough piloting" teachers, (b) "adopting but adapting content" teachers, (c) "adopting but adapting pedagogy" teachers, and (d) "partial piloting" teachers. This suggests that even putatively cooperating teachers are generally disinclined to do what they are told, and emphasises the importance of investigation of IF.

More positively, Crawford, Carpenter, Wilson, Schmeister and McDonald (2012) investigated fidelity and outcomes in a computer-based middle school mathematics curriculum for 485 students and 23 teachers from 11 public middle schools. Total time in intervention, concentration of time in intervention, direct observation of intervention fidelity and pre-test score were all significant, but fidelity to process was nonsignificant. Wolfe, Clements, Sarama, & Spitler (2013) focused on IF over time, examining the sustainability of teachers' implementation fidelity in a prekindergarten mathematics intervention, two years after external support ceased. Teachers continued to demonstrate high levels of fidelity to the underlying curriculum. Kinzie, Whittaker, McGuire, Lee, & Kilday (2015) evaluated the Research on Curriculum Design (RCD) model for pre-kindergarten mathematics and science curricula. Implementation spanned two years and involved iterative development and testing. A final test of the resulting curricula in eight pre-K classrooms yielded high-quality, high-fidelity teacher implementation, with teacher fidelity and curricular dosage predicting students' mathematics learning gains.

However, all but one of these studies focused on a short period of implementation, few reported IF over a longer period and even fewer reported IF indices available as a matter of course without additional effort as in AM. Crucially, none investigated the perceptions of the students, let alone the relevant behaviours of the students, as in the present study.

2.3 Outcome Literature Using STAR Math and Accelerated Maths

There is a good deal of outcome research on STAR Math and AM. The What Works Clearinghouse (2017) has summarized the primary level research (kindergarten through pre-algebra) in its own narrow way. Six studies met the WWC requirements, including 5,206 students in grades 2-9 in 223 classrooms across 27 states. The evidence for impact of AM on the mathematics test scores of students was medium to large. Consequently, there is no need

to demonstrate that AM "works" – that has already been done - although almost all the outcome research has been in the US.

Studies in the US demonstrating the effectiveness of AM include those of Gaeddert (2001), Zumwalt (2001), Spicuzza, Ysseldyke, Lemkuil, Kosciolek, Boys, and Teelucksingh (2001), Ysseldyke, Spicuzza, Kosciolek, and Boys (2003), Ysseldyke, Spicuzza, Kosciolek, Teelucksingh, et al. (2003), Ysseldyke, Betts, Thill, and Hannigan (2004), Springer, Pugalee, and Algozzine (2007), and Stanley (2011).

Teelucksingh, Ysseldyke, Spicuzza, and Ginsburg-Block (2001) studied English Language learners in grades 4-5 in four schools, finding AM students gained twice as much as the controls. Ysseldyke, Tardrew, Betts, Thill, and Hannigan (2004) focused on gifted students. Those who used Accelerated Maths advanced significantly more than those who did not. Ysseldyke and Betts (2010) found that AM was more effective than the following math curricula: enVision Math, Everyday Mathematics, Holt McDougal, Macmillan/McGraw-Hill and Saxon Math.

In Australia, Anamourlis (2001) investigated 250 students from Years 3–7 from five schools throughout Australia over only five months. The teachers did not receive training and. AM and control groups showed similar gains in number, but the AM group showed very large relative gains in areas of maths outside number. In the UK, Knock (2005) used AM in a daily 30-minute lunch-time maths club deploying AM. The AM group improved on attainment tests three times more than the comparison group. Rudd and Wade (2006) matched 14 schools implementing AM with seven comparison schools. Comparisons over an eight-month period showed good results for AM in secondary schools, but less god results in primary schools.

The extent to which different components of implementation contribute to these outcomes is another story, which we shall now explore.

2.4 Outcome and Implementation Fidelity in Accelerated Maths

AM offers a novel way of assessing IF, by abstracting indices of student responsiveness directly through computers (see Methodology section below for a fuller description of AM). However, relatively little of the previous literature on IF of AM has adopted this approach.

An early study in Germany examined AM with 22 fourth-, fifth-, and sixth-grade classrooms in 14 schools, matched with an approximately equal number of same-school, same-grade control classrooms that used their regular instructional methods (Lehman & Seeber, 2005). Fifth-grade students using AM increased twice as much as the control group, but in grades four and six, AM and control students experienced similar levels of growth. Classrooms in which the AM program was used very extensively achieved the largest gains.

All other studies were in the US. Holmes, Brown, and Algozzine (2006) examined the effectiveness of AM with 2,287 students from four elementary schools (two rural, two urban). One school in each area was either a high or low implementer of AM. Students in the two high-implementing schools outperformed students in the two low-implementing comparison schools overall (effect size [ES]=0.65) and in math (ES=0.75). The impact of randomly assigned Accelerated Maths with nearly 2,000 elementary students from eight schools and 100 classrooms in eight states was examined by Ysseldyke and Bolt (2007). Students whose teachers used AM *as intended* demonstrated greater gains than students with limited or no implementation.

Investigating 11 AM schools and 11 matched control schools, Nunnery and Ross (2007) found that both elementary and middle school students benefited from using AM, especially at high-implementing schools. Bolt, Ysseldyke, and Patterson (2010) found students of teachers who implemented AM with greater fidelity experienced higher math gains on standardized assessments than other students. Burns, Klingbeil, and Ysseldyke

(2010) studied 360 randomly selected schools in four states to compare AM with controls. AM students scored significantly more proficient on their states' high-stakes tests than the control group. An achievement gap existing in control schools between white and ethnic-minority populations did not exist at the treatment schools.

A sample of over 18,000 English language learners (ELs) and Native English Speakers (NESs) were studied by Lekwa (2012). Implementation of AM was a strong predictor of math skill growth for both ELs and NESs. Walker Driesel (2013) examined pre and post-test scores on the STAR Math standardized test in relation to amount of classroom time dedicated to AM instruction. There was a strong correlation between student performance and amount of time for AM. Lambert, Algozzine, and McGee (2014) categorized AM treatment classes in grades 2–5 at three Midwestern elementary schools into high- or low-implementation groups. Growth for the high-implementation group was significantly higher than for the low-implementation group, although both groups did better than non-AM controls.

However, none of these studies directly used computer-based student-led indices of IF. An exception was the quasi-experimental study of Ysseldyke and Tardrew (2007), who investigated 2,200 students from 47 schools in 24 US states. Students using AM in grades 3–10 achieved math gains from 7-18 percentile points higher than comparison students. In every grade and in Title I and free lunch programs, students in AM classes outperformed students not using the program. Low-, middle-, and high-achieving students showed consistent rates of gain. Importantly, students who followed AM best practice recommendations by scoring higher than 85% correct and completing more subskills made even greater gains. This study was most like the present study.

The linkage between student behaviour and attainment in maths is important in the light of the theory of situated learning (Lave & Wenger, 1990). This posits that learning

activities need to be presented in authentic contexts - settings and situations that would normally involve that learning. If the learning is thus contextualized, the student is in the best position to monitor if that learning is effective. However, given that student meta-cognition may not be well-developed, examination of student learning behaviours rather than student perceptions is highly advantageous.

3.0 Research Questions

1. Is Accelerated Math well implemented?
2. Is better performance in student-indicated implementation quality of Accelerated Maths associated with better mathematics outcomes on STAR Math?
3. Do socio-economic status, age, mathematics ability and gender influence these findings?

4.0 Method

4.1 Sample

The sample comprised all students in the UK for whom STAR results were available for the academic year (n = 20,103 in 75 elementary and 73 high schools). Of these, the number of students for whom AM data were available was n=6,285 (further details are given in Results under Implementation). The grades represented ranged from 2 through 13 and were approximately normally distributed, although grade 7 was under-represented and grade 8 was over-represented. The sample was generally representative of the whole of the UK, although Scotland was poorly represented (see the last section of the Results). However, the number of students for each analysis was generally large and is noted in the text.

4.2 Measures

4.2.1 Attainment - STAR Math. STAR Math is a computerized standardized (norm-referenced) adaptive item-banked math test. It has a mathematics question with multiple-choice answers on each page. It is standardized, i.e. any student's test responses are

compared with the responses of many students of that age. The test is adaptive, i.e. it responds to the performance of each individual student. If the pupil succeeds on a question, 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 at the same time. 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 Math has test-retest reliability of 0.93 in a US national sample of more than nine million tests. It also has internal consistency reliability of 0.97. More than 400 concurrent and predictive validity studies (correlations with many other measures of mathematics achievement) have been collected for STAR Math, involving 400,000 students. The average validity correlations range from 0.55 to 0.80. Correlations in that range are considered moderate to strong (Renaissance Learning, 2013).

4.2.2 Implementation fidelity - Accelerated Maths. AM is a personalized practice and progress-monitoring system that customizes math practice assignments for students, gives instant feedback to the student and helps teachers accurately and efficiently monitor pupil progress in quantity, difficulty and mastery of mathematics skills. The program does not provide instruction - it does however provide practice in carefully differentiated skills for the student, a system of scoring and monitoring, and a system of feedback to the student and the teacher. Each student is assigned by the teacher to a series of practice activities on math objectives, initially based on the student's entry score on STAR Math. AM automatically scores student work, and students and teachers can view feedback reports that describe the nature of the performance. The student receiving this feedback can then reflect on how he/she should respond or behave better on the next task in order to improve this feedback. AM then assigns new activities which are adaptive to the performance of the student on the previous

task. If the student performed exceptionally well, a considerably harder task of the same type will be delivered. If moderately well, a somewhat harder task. If not at all well, a task at the same level or an easier task. After reviewing student progress, teachers can adjust instruction for the entire class, for small groups of students struggling with similar objectives, or for individual students.

AM currently includes content for grades K–8 in Algebra I, Geometry, and Algebra II. The equipment required is a class computer, printer and an Optical Mark Reader. Since students do their work on paper at their desks, and not at the computer, one set of equipment can serve the entire class.

Definitions of terms used in this paper are now offered.

4.3 Definitions

4.3.1 Achievement. 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 1-99. 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.

4.3.2 Implementation. All of the following metrics are driven by student behaviour. For each assessment in any activity, Average Percent Correct (APC) is the percent of correctness of the student's answers to the questions (in this case aggregated over a year).

"Diagnostics" is an assessment of previous student work on AM to gauge overall knowledge and identify any gaps in skills or other problems. The APC on Diagnostics has a recommended level of performance of >=85% correct or higher.

Each objective has three sets of mastery criteria: Practice, Test, and Review.

"Practice" is the number of practice or exercise problems a student must answer correctly to show they are ready for testing. It checks whether the student is practicing the right skills at the appropriate level, time and pace. The APC on Practice has a recommended level of performance of >=75% correct or higher.

"Test" is the number of test problems a student must answer correctly to master an objective. It assesses the students' level of mastery of the objective, once they have practiced skills sufficiently. The recommended level of performance is >=85% correct or higher.

Review problems appear on practices starting two weeks after students master an objective, and are designed to sustain student ability to answer questions on material previously learned. "Review" is the number of review problems the student must answer correctly to complete work on a past objective. The recommended level of performance is >=80% correct or higher.

"Objectives Mastered" is a count of mastery of the Objectives (concepts and subskills) each student has been learning.

4.4 Data Analysis

Initially descriptive statistics were used to illuminate the data. Non-parametric correlation was carried out to examine relationships between variables. Given the size of the database, we opted not to use inferential statistics, as these are considerably affected by

sample size. Instead we chose to use Effect Sizes (ES) as indicators of magnitude of effect. Cohen's delta effect sizes were calculated to examine the importance of differences between variables. Effect Sizes of .10 were characterized as "very small" (Sawilowsky, 2009). Effect Sizes of .20 were characterized as "small", those of .50 as "moderate" and those of .80 as "large" (Cohen, 1988). Effect Sizes of 1.20 were characterized as "very large" and those of 2.0 as "huge" (Sawilowsky, 2009). In the interests of clarity and transparency, break points between these indices were added by the present author: 0.10 between .01 and .20, .35 between .20 and .50, .65 between .50 and .80, 1.00 between .80 and 1.20, and 1.60 between 1.20 and 2.0.

5.0 Results

In these Results, we will first examine attainment on the STAR Math test, reporting socioeconomic status differences. Then we will examine AM implementation variables and attainment. Further analysis of gender, primary-secondary status (related to grade), socio-economic status and math ability differences in implementation follow.

Attainment Data

On the standardized attainment test STAR Math, a total of n=19,841 students had SGP scores with mean 52.10 (sd 28.88). This suggested that students taking STAR Math performed overall at above the average level. Students receiving Free Lunch or Not (as an indicator of socio-economic status) were recorded for 19,283 cases (96%). Of these, 1,095 had Free Lunch, while 18,188 did not. Free Lunch students scored an average of 52.71 (sd=29.64), while Not Free Lunch students scored 52.14 (sd=29.90). Thus, Free Lunch students did better than Not Free Lunch students, but the ES was only .02 (very small). Nonetheless, the usual expectation that low-SES students will do poorly was not supported on the STAR Math test.

Implementation Data

The number of students yielding implementation data on APC for Diagnostics, APC for Practice, APC for Test, APC for Review, and Objectives Mastered was much smaller than the number yielding attainment data, but still substantial (n=6,285). There were slightly more boys than girls, but there was no difference between genders in attainment or implementation – the results were identical. The number of students having APC_Practice scores was n=6,285, the largest number. Other implementation scores were based on n=5677 for APC_Test, n=5530 for Objectives Mastered, n=4803 for APC_Review, and n=2370 for APC_Diagnostics. However, the average SGP for the smaller sample was very similar to that for the larger sample (SGP mean = 53.16, sd = 29.48). The ES comparing large to small samples for SGP was .04 (very small).

Non-parametric correlations (Spearman's) between attainment and implementation variables were undertaken. Correlations between attainment and implementation were modest. The highest correlation was between APC_Practice and attainment - .24. The next highest was Objectives Mastered at .12 (SGP). These were small and accounted for relatively little of the variance.

However, the implementation variables showed relatively high correlation with each other. APC_Practice correlated at .51 with Objectives Mastered, .38 with APC_Test, .51 with APC_Review, and .54 with APC_Diagnostics. Objectives Mastered correlated at .36 with APC_Test, .30 with APC_Review and .43 with APC_Diagnostics. All of these correlations were statistically significant on account of the relatively high numbers in the sample.

High Quality Implementation

Renaissance Learning recommends criteria indicating high quality implementation. As noted earlier, these are APC_Diagnostics>=85%, APC_Practice >= 75%,

APC_Test>=85%, and APC_Review>=80%. Objectives Mastered does not lend itself to such recommendations since different objectives are not at the same level of difficulty, although obviously more is better at any level.

Only 403 students (6.4%) had APC_Practice scores >=75%. Their mean was .82 (sd=.06), with SGP 64.60 (sd=28.32). APC_Practice<75% (n=5882) had a mean of .40 (sd=.16), with SGP of 52.44 (sd=29.41). The APC_Practice <> mean difference was very substantial with ES 3.88 (huge). The SGP mean difference was quite substantial at ES .35 (moderate). Thus, there was considerable evidence validating this high-quality implementation criterion, although unfortunately only a small percentage of students came into this category.

Only 492 students (8.7%) had APC_Test scores >=85%. Their mean was .91 (sd=.05), with SGP Gain 59.46 (sd=29.52). APC_Test <85% (n=5184) had a mean of .47 (sd=.24), with SGP Gain 53.32 (sd=29.29). The APC_Test <> mean difference was very substantial with ES 3.10 (huge). The SGP mean difference was fairly substantial at ES .21 (small). Thus, there was some evidence validating this high-quality implementation criterion, although unfortunately only a small percentage of students came into this category.

Only 552 students (11.5%) had APC_Review scores>=80%. Their mean was .89 (sd=.07), with SGP 60.57 (sd=29.13). APC_Review <80% scores (n= 4251) had a mean of .49 (sd=.21), with SGP 53.06 (sd=29.17). The APC_Review <> mean difference was very substantial with ES 2.83 (huge). The SGP mean difference was at ES .26 (small). Thus, there was some evidence validating this high-quality implementation criterion, although unfortunately only a small percentage of students came into this category.

Only 259 students (11.0%) had APC_Diagnostics scores>=85%. Their mean was .94 (sd=.05), with SGP Gain 60.10 (sd=30.25). APC_Diagnostics <85% scores (n=2119) had a mean of .331 (sd=.26), with SGP Gain 50.88 (sd=28.91). The APC_Diagnostics <> mean

difference was very substantial with ES 3.93 (huge). The SGP mean difference was fairly substantial at ES .31 (small). Thus, there was considerable evidence validating this high-quality implementation criterion, although unfortunately only a small percentage of students came into this category.

Thus, for all these four indices of IF, the differences between above recommended levels of implementation and below recommended levels of implementation were very large, with all the Effect Sizes being "huge". In addition, high implementers tended to be higher on attainment.

Grade Level in Relation to Implementation and Attainment

We needed to examine whether the relationships between attainment and implementation were moderated by Grade, since it was possible that there was considerable variation between Grades, with positive and negative results cancelling each other out to make it appear as if there were no difference. For SGP, Primary was ahead of Secondary (ES .15 – small). For APC_Diagnostics, Primary was ahead of Secondary (ES .55 – moderate). For APC_Practice, Primary was ahead of Secondary (ES .40 – moderate). For APC_Tests, Secondary was ahead of Primary (ES .06, very small). For APC_Review, Secondary was ahead of Primary (ES .07 – very small). For Objectives Mastered, Primary was ahead of Secondary (ES .64 – moderate). Thus, the Primary sector appeared to show the largest effects - on Diagnostics, Practice and Objectives Mastered.

However, inspection of attainment and implementation by Grade revealed that the Primary/Secondary distinction was masking more complex differences between Grades (see Table 1). As with the whole attainment sample, Years 6 and 7 showed the best attainment results - one year in primary and the other in secondary. Performance was less good before and after this. Diagnostics were highest in Years 4 and 7. Practice scores were highest in Year 5. Test scores were highest in Years 6 and 7-8. Review scores were highest in Years 5,

6 and 7. Objectives Mastered was high in Years 4, 5 and 6. In general the highest scores were in upper primary and lower secondary.

INSERT TABLE 1 ABOUT HERE

Socio-economic Status in Relation to Implementation and Attainment

We needed to check whether the relationship between implementation and attainment was affected in any way by Free Lunch or Not (socioeconomic status), since it was possible that it was higher for low socio-economic status students. Table 2 shows that only 554 students (9%) yielding attainment and implementation data had Free Lunch. Free Lunch students were a little below Not Free Lunch students on attainment, but ES for SGP was .01 (very small). On Objectives Mastered, Free Lunch students were at almost the same level as Not Free Lunch students (ES<.01 – very small). Not Free Lunch students were ahead on three of the four APC measures (ES APC_Tests .12 small, ES APC_Practice .15 small, ES APC_Diagnostics .32 small). However. Free Lunch students were ahead on APC_Review – ES .04 very small). Overall there was little difference between Free Lunch and Not Free Lunch students and we can conclude that socio-economic status was not a factor in implementation or attainment.

INSERT TABLE 2 ABOUT HERE

High/Low Ability in Relation to Attainment and Implementation

Similarly, we needed to examine the relationship of ability to Implementation and Attainment, since it was possible that the programme had a differential effect, perhaps favouring the low ability learner. In relation to ability in mathematics, we could assume that SGP would indicate this with a mean of 100 and a standard deviation of 15. Thus, high achievers could be seen as students with SGP at or above 115, while low achievers could be seen as those with SGP at or below 85. We tried to compare high with low achievers on this

basis but found that there were no high achievers (under this definition). Consequently, we compared low achievers with the average for all students (Table 3).

INSERT TABLE 3 ABOUT HERE

Unsurprisingly, average achievers did better on SGP (ES .35 - moderate). However, there was little difference between average and low achievers in implementation: APC_Diagnostics ES .06, APC_Practice .05, APC_Tests ES .06, APC_Review ES .04, and Objectives Mastered ES .06 – all very small.

6.0 Discussion

6.1 Summary

The main finding of this study was that implementation of AM was very poor in the UK, with only 6.4% - 11.5% percent of pupils scoring at or above the recommended levels of implementation. Implementation quality had a modest correlation with attainment, perhaps because the levels of implementation were so low. However, high implementation was associated with high attainment. Socio-economic status did not seem to have any effect on attainment or implementation. Age of pupil did have some effect on implementation, varying according to the index of implementation in question, but in general upper primary and lower secondary did best. Neither gender nor mathematics ability had any significant influence on attainment or implementation. However, AM pupils still performed above average on the maths attainment.

In more detail, in attainment for the larger sample (n=20,103), the average STAR Math SGP was 52.10, suggesting overall student performance was above average, perhaps surprising for those who conceive of AM as a remedial programme. Males and females had identical attainment, so there was no gender effect with boys outperforming girls. Low socio-economic students did slightly better than the rest of the sample, although effect sizes were

small. Nonetheless, the usual expectation that low-SES students will do poorly was not supported on the test.

Regarding attainment *together with* implementation indices, a much smaller sample was available (n=6,285). SGP levels did not differ from the large sample and correlated highly. However, correlations between attainment and implementation indices were much more modest (maximum .24, accounting for relatively little of the variance). APC_Practice correlated best with attainment. However, implementation variables correlated quite well with each other (maximum .54). All of these correlations were statistically significant on account of the relatively high numbers in the sample. Again, there was no difference between genders.

Considering high quality implementation, recommended levels are APC_Diagnostics>=85%, APC_Practice >=75%, APC_Tests>=85%, and APC_Review>=80%. Only 259 (11%) of students achieved this level for APC_Diagnostics, and their ES for implementation was 3.93 (huge) and for SGP .31 (small). Only 403 (6%) of students achieved this level for APC_Practice, and their ES for implementation was 3.88 (huge) and for SGP .35 (moderate). Only 492 (10%) of students achieved this level for APC_Tests, and their ES for implementation was 3.10 (huge) and for SGP .21 (small). Only 552 (11%) of students achieved this level for APC_Review, and their ES for implementation was 2.83 (huge) and for SGP .26 (small). Thus, there was good evidence of a positive relationship between high implementation and high attainment. For all four indices of IF, the differences between above recommended levels of implementation and below recommended levels of implementation were very large, with all the Effect Sizes being “huge”.

Examining differences between grade levels (Years) to investigate whether this pattern was consistent across Years, the Primary sector appeared to show the largest effects - on Diagnostics, Practice and Objectives Mastered. However, analysing by individual year

showed in general the highest scores were in upper primary and lower secondary. Overall there was little difference between Free Lunch and Not Free Lunch students and we can conclude that socio-economic status was not a factor in effectiveness. Likewise, there was little difference between average and low achievers in implementation – all ESs were very small.

6.2 Connection to Previous Literature

As noted in the review of previous literature, ten studies found implementation quality related to attainment (and none did not). In some cases, this was over long periods, such as two years and five years. All studies were in the US except for one in Germany. However, assessment of quality of implementation was generally not in terms of the implementation indices used here. An exception to this is the study of Ysseldyke and Tardrew (2007), who noted that students who closely followed AM implementation recommendations by scoring higher than 85% correct and completing more subskills (Objectives Mastered) made the largest gains. This study is most like the present study, except it took place several years ago in a different country and did not use all of the implementation indices. The present study is from a different country and uses all five implementation indices. This study confirms that Objectives Mastered is important, but adds Diagnostics, Practice, Test and Review as further key indicators of implementation.

6.3 Limitations and Strengths

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 a de-emphasis on statistical significance. The sample was representative of grades from 2-13 and of the United Kingdom, except for Scotland. SGP tends to under-estimate schools in socio-economically disadvantaged areas and over-estimate schools in advantaged areas, so when interpreting the

tables some flexibility is required. The low level of implementation possibly affected the association between implementation and attainment.

6.4 Implications for Practice, Policy and Future Research

6.4.1 Practice. Teachers and students should strive to raise implementation levels of what appear to be the major determinants of higher outcomes - APC_Diagnostics, APC_Practice, APC-Tests, APC_Review, and Objectives Mastered – to the recommended levels. 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 AM in their schools, they should carefully consider the evidence on these key indicators of IF as well as the level of student attainment outcomes and strive to increase them to recommended levels. Students should also respond more thoughtfully to the feedback they receive from the AM system.

6.5.2 Policy. Policy-makers (including school inspectors) at local and national level should carefully consider the evidence on these key indicators of IF as well as the level of attainment. Policy-makers need to be sharply aware that 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 should consider providing relevant professional development opportunities to teachers and schools.

6.5.3 Future research. *S*hould studies similar to this be undertaken, in whatever country, it would be useful to investigate all the implementation variables found effective here and their relationship to attainment. A further comparative study of indirect, direct and computerized methods of establishing IF in mathematics with the same pupils would be highly valuable. As in this study the level of adequate implementation was so low, conducting further analysis of the relationship between implementation and attainment with these data would be of limited usefulness. If the level of implementation was higher,

regression analyses linking implementation with attainment and allowing the prediction of attainment from level of implementation would be possible.

7.0 Conclusion

Thus, in relation to the research questions, we found that AM was not well implemented (RQ1). We found that better performance in the implementation quality of Accelerated Math was modestly correlated with better mathematics outcomes on STAR Math, but high implementation was associated with high attainment (RQ2). Overall, on average AM students performed at above average levels. We found that socio-economic status had no effect on either attainment or implementation (RQ3), which was a surprising finding, suggesting that STAR and AM might be culture-fair. Age of pupil did have some effect on attainment and implementation, depending on the implementation index used, but in general upper primary and lower secondary did best (RQ3). Mathematics ability had no effect on attainment or implementation, so there was no evidence that less able or more able pupils fared better on AM (RQ3). Age and gender showed no effects (RQ3).

Computerized student-response measures of IF might have problems in reliably predicting pupil outcome, just like teacher opinion and direct observation. It is suggested that future research needs to triangulate indirect, direct and computerized 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.

Computerized student response measures are not yet available in many areas of the curriculum. Computerized methods of assessing teacher behaviour also seem to be some way in the future. Nonetheless, a much larger portion of research resource needs to be devoted to establishing satisfactory multi-component IF measures. Researchers and research organizations interested in evidence-based interventions need to give much closer attention to the issue of IF.

References

Anamourlis, A. (2001). *The impact of the Accelerated Maths pilot program in Australia.* Boxhill: Renaissance Learning Australia.

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 { HYPERLINK "http://www.nj.gov/education/njsmart/performance/SGP_Technical_Overview.pdf%20%5b14" } June 2016].

Bolt, D., Ysseldyke, J., & Patterson, M. (2010). Students, teachers, and schools as sources of variability, integrity, and sustainability in implementing progress monitoring. *School Psychology Review, 39*(4), 612–630.

Burns, M. K., Klingbeil, D. A., & Ysseldyke, J. (2010). The effects of technology enhanced formative evaluation on student performance on state accountability math tests. *Psychology in the Schools, 47*(6), 582–591.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York and London: Routledge. ISBN 1-134-74270-3.

Crawford, L., Carpenter, D. M., Wilson, M. T., Schmeister, M., & McDonald, M. (2012). Testing the relation between fidelity of implementation and student outcomes in math. *Assessment for Effective Intervention, 37*(4), 224-235.

Gaeddert, T. (2001). *Using Accelerated Math to enhance student achievement in high school mathematics courses.* Unpublished master's thesis. Friends University, Wichita, Kansas. ERIC Number: ED463177.

Holstein, K. A. (2012). *A characterization of teachers' implementations of a mathematical decision-making curriculum*. ProQuest LLC, Ph.D. Dissertation, North Carolina State University. ERIC Number: ED550966.

Kinzie, M. B., Whittaker, J. V., McGuire, P., Lee, Y. J., & Kilday, C. (2015). Research on curricular development for pre-kindergarten mathematics and science. *Teachers College Record, 117*(7), 136-157.

Holmes, C. T., Brown, C. L., & Algozzine, B. (2006). Promoting academic success for all students. *Academic Exchange Quarterly, 10*(3), 141–147.

Knock, D. J. (2005). *Nottingham pupils improve mathematics achievement with Accelerated Maths.* Wisconsin Rapids, WI: Renaissance Learning.

Lambert, R., Algozzine, B., & McGee, J. (2014). Effects of progress monitoring on math performance of at-risk students. *British Journal of Education, Society and Behavioural Science, 4*(4), 527–540. Retrieved from { HYPERLINK "http://www.journalrepository.org/media/journals/BJESBS_21/2014/Jan/Lambert442013BJESBS7259_1.pdf" }

Lave, J., & Wenger, E. (1990). *Situated learning: Legitimate peripheral participation.* Cambridge, UK: Cambridge University Press.

Lehman, R. H., & Seeber, S. (2005). *Accelerated Mathematics in grades 4-6.* Wisconsin Rapids, WI: Renaissance Learning.

Lekwa, A. J. (2012). *Technology-enhanced formative assessment in mathematics for English language learners* (Unpublished doctoral dissertation). University of Minnesota, Minneapolis. ERIC Number: ED551876.

Nunnery, J. A., & Ross, S. M. (2007). The effects of the School Renaissance program on student achievement in reading and mathematics. *Research in the Schools, 14*(1), 40–59.

Randel, B., Apthorp, H., Beesley, A. D., Clark, T. F., & Wang, X. (2016). Impacts of professional development in classroom assessment on teacher and student outcomes. *Journal of Educational Research, 109*(5), 491-502.

Renaissance Learning (2013). *The research foundation for Star Assessments.* Wisconsin Rapids, WI: Renaissance Learning.

Rudd, P., & Wade, P. (2006). *Evaluation of Renaissance Learning mathematics and reading programs in UK Specialist and feeder schools.* Slough: National Foundation for Educational Research.

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.

Sawilowsky, S (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, 8(2), 467–474. { HYPERLINK "http://digitalcommons.wayne.edu/jmasm/vol8/iss2/26/" }

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.

Spicuzza, R., Ysseldyke, J., Lemkuil, A., Kosciolek, S., Boys, C., & Teelucksingh, E. (2001). Effects of curriculum-based monitoring on classroom instruction and math achievement. *Journal of School Psychology, 39*(6), 521–542.

Springer, R. M., Pugalee, D., & Algozzine, B. (2007). Improving mathematics skills of high school students. *The Clearing House, 81*(1), 37–44.

Stanley, A. M. (2011). *Accelerated Mathematics and high-ability students' math achievement in grades three and four.* ProQuest LLC, Ed.D. Dissertation, East Tennessee State University. ERIC Number: ED532232.

Teelucksingh, E., Ysseldyke, J., Spicuzza, R., & Ginsburg-Block, M. (2001). *Enhancing the learning of English language learners: Consultation and a curriculum-based monitoring system.* Minneapolis, MN: University of Minnesota, National Center for Educational Outcomes.

Topping, K. J. (2017). Implementation fidelity in computerised assessment of book reading. *Computers and Education, 116*, 176-190. doi: 0.1016/j.compedu.2017.09.009.

Topping, K. J. (2018). Implementation fidelity and pupil achievement in book reading: Variation between regions, local authorities and schools. *Research Papers in Education, 33*(5), 620-641. doi: 10.1080/02671522.2017.1329340.

Walker Driesel, D. (2013). *Mathematics interventions: A correlational study of the relationship between level of implementation of the Accelerated Math program and student achievement*. ProQuest LLC, Ed.D. Dissertation, Liberty University. ERIC Number: ED564886.

What Works Clearinghouse (2017). *Accelerated Math. Primary Mathematics*. Washington, DC: Institute of Education Sciences.

Wolfe, C. B., Clements, D. H., Sarama, J., & Spitler, M. E. (2013). *Sustainability of fidelity of implementation over time in the context of a prekindergarten mathematics curriculum and professional development scale-up intervention.* Washington, DC: Society for Research on Educational Effectiveness. ERIC Number: ED564101. Available https://files.eric.ed.gov/fulltext/ED564101.pdf [12 June 2018].

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 [16 June 2016].

Ysseldyke, J., & Betts, J. (2010). *Progress monitoring, implementation integrity, and guided practice across multiple math curricula.* Presentation at Council for Exceptional Children 2010 Annual Convention & Expo, Nashville, TN.

Ysseldyke, J., Betts, J., Thill, T., & Hannigan, E. (2004). Use of an instructional management system to improve mathematics skills for students in Title I programs. *Preventing School Failure, 48*(4), 10–14.

Ysseldyke, J., & Bolt, D. (2007). Effect of technology-enhanced continuous progress monitoring on math achievement. *School Psychology Review, 36*(3), 453–467.

Ysseldyke, J., Spicuzza, R., Kosciolek, S., & Boys, C. (2003). Effects of a learning information system on mathematics achievement and classroom structure. *Journal of Educational Research, 96*(3), 163–173.

Ysseldyke, J., Spicuzza, R., Kosciolek, S., Teelucksingh, E., Boys, C., & Lemkuil, A. (2003). Using a curriculum-based instructional management system to enhance math achievement in urban schools. *Journal of Education for Students Placed at Risk, 8*(2), 247–265.

Ysseldyke, J., & Tardrew, S. (2007). Use of a progress-monitoring system to enable teachers to differentiate mathematics instruction. *Journal of Applied School Psychology, 24*(1), 1–28.

Ysseldyke, J., Tardrew, S., Betts, J., Thill, T., & Hannigan, E. (2004). Use of an instructional management system to enhance math instruction of gifted and talented students. *Journal for the Education of the Gifted, 27*(4), 293-319.

Zumwalt, D. B. (2001). *The effectiveness of computer-aided instruction in eighth-grade pre-algebra classrooms in Idaho* (Unpublished PhD dissertation). Idaho State University,

Idaho. Retrieved June 12, 2018 from { HYPERLINK

"https://www.learntechlib.org/p/117250/" }.