RAND EUROPE

Evaluation of a Technology-based Intervention for Reading in UK Classroom Settings

Hannes Jarke, Miriam Broeks, Sashka Dimova, Eleftheria Iakovidou, George Thompson, Sonia Ilie, Alex Sutherland

Microsoft provided the funding that supported the in-classroom implementation and evaluation of the technology-based intervention that made use of Immersive Reader.

The authors would like to acknowledge the efforts of *Achievement for All*, and in particular Marius Frank, Nicole Ponsford, Professor Sonia Blandford, and colleagues, for leading the implementation of Immersive Reader and for facilitating data collection from the pupils, teachers, head teachers, and schools participating in the study.

The authors thank Elta Smith and Axelle Devaux from RAND Europe for their reviews of this document and providing impartial feedback.

For more information on this publication, visit www.rand.org/t/RR4208

Published by the RAND Corporation, Santa Monica, Calif., and Cambridge, UK

RAND Europe is a not-for-profit research organisation that helps to improve policy and decision making through research and analysis. RAND's publications do not necessarily reflect the opinions of its research clients and sponsors.

Executive summary

The use of digital teaching aids to improve reading is of interest to teachers around the world. RAND Europe evaluated the implementation of a technology-based intervention that made use of the Microsoft Immersive Reader (IR) in UK classrooms for eight- and nine-year-old pupils. This is the first large-scale randomised control trial (RCT) of this intervention, following promising findings related to the use of IR from a small-scale study in the United States. The study randomly allocated Year 3 and Year 4 classes in 20 schools to either an intervention condition (receiving the intervention including IR) or a control condition (business as usual). Using the standardised Progress in Reading Assessment (PiRA) test, this study compared reading scores for pupils in the randomly allocated IR condition to class-level reading outcomes for pupils in the control condition.

The retention of schools in the study proved challenging, so that by the end of the trial, whereby of the 80 initially-recruited classes, only 54 completed the final reading assessment, limiting the internal validity of the analysis. Additionally, teachers' engagement with the data collection processes was also limited, so that the perspectives outlined in this report do not emerge from a representative sample of study participants, and therefore cannot be assumed to be universally applicable. Lastly, technical difficulties in relation to the reading assessment platform led to a study re-design whereby only one of the three initial hypotheses was explored in the pupil data.

Given the low fidelity of implementation practices as described below, the infrequent use of the intervention in some schools, as well as the high trial attrition rate, it is not possible to ascertain if the intervention would have influenced the students' reading attainment had it been implemented as intended. This means that on average, pupils in the intervention classes performed as well as pupils in control classes.

Separately, the evaluation of the process of implementing the intervention revealed substantial inconsistencies and deviations from the intended implementation plan, explored extensively in this report. Teachers reported a general openness towards, and interest in, the technology itself, but also reported feeling overwhelmed with technical issues, time and resource constraints (such as not enough tablets or computers). Teachers also reported using different features of IR with different frequencies across a variety of classroom contexts (e.g. with pupils with special educational needs or with low reading attainment). This suggests that the intervention was not implemented as intended in many of the participating schools, with substantial variation between teachers' reported practices.

This added a further challenge to the internal validity of the study, and any future evaluations of similar interventions, including IR, would benefit from the careful monitoring of implementation fidelity and processes.

From the perspective of on-the-ground implementation, therefore, any future implementation and testing of an intervention including IR should consider three areas. Firstly, implementation clarity should be established early on (so that schools and teachers are aware of all the implications of participation). Further, implementation should be consistent across participating schools, classes and teachers, including a measure of implementation fidelity (to ensure that any effectiveness study is concerned with one consistent intervention that adheres to the design of the intervention). Lastly, support should be given to schools and teachers to deal with technical difficulties that may impede the implementation of the intervention. With implementation issues addressed, future trials can reassess the impact intervention including IR may have on pupils' reading outcomes.

Introduction

Immersive Reader

Immersive Reader (IR) is a tool that aims to enhance reading instruction and learning in schools. It implements techniques to improve reading and writing for all pupils, and has features to customise reading experience by, for example, highlighting text while reading it out loud, breaking text into syllables and increasing spaces between words. IR can be used as a teaching aid to help teachers support their pupils' learning and encourage independent learning. By improving pupil learning and promoting reading independence, IR may lead to improvements in pupil attainment. IR was first developed in a 2015 hackathon hosted by Microsoft[1] and became available as a free add-on to the Microsoft digital note taking app OneNote. It has since been incorporated in many products and services across platforms.

Previous evidence from exploratory research conducted in the United States showed that Microsoft Learning Tools such as IR can support reading and writing, especially for struggling readers. For instance, a previous small-scale study conducted with eight teachers in Washington State yielded promising results.[2] Four of the teachers worked with primary school pupils who used IR for one semester and showed greater gains in reading comprehension compared to a historical cohort of students from the same school. Teachers in the study also highlighted positive perceived impact of their use of the learning tools in their classroom. Previous research has also suggested, however, that providing teachers with technology was not sufficient to lead to changes in practices in the classroom and to improvements in learning

1 A hackathon is a time-constrained event in which programmers and related experts try to collaboratively design a software project. During the 2015 Microsoft hackathon, IR was one of the winning projects. https://www.windowscentral.com/how-microsofts-immersive-reader-tackles-dyslexia-head-and-wins

2 McKnight, K. 2017. *Leveling the Playing Field with Microsoft Learning Tools*. Centre for Evaluation & Study of Educational Equity, RTI International.

for pupils.[3,4,5,6] Additionally, training in the use of technology reportedly remains one of the areas in which teachers have the greatest need for continued professional development. [7,8,9]

Evaluating the effectiveness of IR

While results from the US study appear promising, no systematic investigation of IR in the UK context has been conducted. As a result, Microsoft granted funding to the delivery team Achievement for All (AfA) to deliver the IR intervention outlined below, with the aim of assessing its impact on pupils' reading outcomes. AfA aimed to implement IR in approximately 80 classrooms across 20 different primary schools in England. RAND Europe (the evaluation team) was subcontracted as the independent evaluator by AfA to design and undertake a two-arm randomised controlled trial (RCT). The RCT is an experimental research design that allows for causal inferences about the impact of an intervention to be drawn. A two-arm trial involves random allocation of participants to one intervention and one control condition.

This is the first evaluation of IR that uses an RCT design. This study focuses on the impact of the intervention including IR on reading attainment of pupils aged 8 and 9 (Year 3 and Year 4 classes) in the UK. This RCT was designed to include an implementation and process evaluation component (IPE), which provides a detailed understanding of the realities of using IR and how it was perceived by teachers and pupils.

Roles

AfA oversaw data collection and aggregation, while RAND Europe had overall analytical responsibility and led the implementation and process evaluation activities. AfA's own findings are included in a delivery team report which is summarised in Box 1 at the conclusion of this report.

The IR intervention in this study

The IR intervention package evaluated in this study consisted of the elements below. The implementation and process evaluation section provides further details as to how the implemented IR intervention differed from the initial plan.

1. Training of IR coaches: Coaches all had primary education experience or were literacy experts. All were experienced school coaches. They were supposed to be trained in IR by AfA in April and May 2018 (prior to the start of school implementation). Delays in in project sign off caused this to happen in June 2018.

---

3 OECD. 2015. *Students, Computers and Learning. Making the Connection.* Paris: OECD Publishing. doi: 10.1787/9789264239555-en

4 OECD. 2012. *Connected Minds: Technology and Today's Learners*,. Paris: OECD Publishing. doi: 10.1787/9789264111011-en.

5 Kirschner, Paul A. 2015. 'Do we need teachers as designers of technology enhanced learning?."?' *Instructional Science*, 43(2): 309–22.

6 Vrasidas, Charalambos. 2015. 'The rhetoric of reform and teachers' use of ICT."'. *British Journal of Educational Technology*, 46(2): 370–80.

7 Barrera-Pedemonte, Fabian. 2016. 'High-Quality teacher professional Teacher Professional Development and classroom teaching. Classroom Teaching Practices: Evidence from TALIS 2013."'. *OECD Education Working Papers*, No. 141, Paris: OECD Publishing.

8 OECD. 2015. Teaching with technology, Teaching in Focus, No. 12,. Paris: OECD Publishing. doi: 10.1787/5jrxnhpp6p8v-en

9 OECD. 2014. *New Insights from TALIS 2013: Teaching and Learning in Primary and Upper Secondary Education.* Paris: OECD Publishing. doi: 10.1787/9789264226319-en

2. Assignment of teachers to schools: Once assigned, the coaches met with senior leaders to ensure
    - that the technical IR deployment within the school was ready and compatible with the aims of the project;
    - arrangements for administering baseline and follow-up online testing were in place;
    - follow-up sessions to work with treatment class teachers were planned for the coming academic year.
3. Introduction workshop for teachers and school Senior Leadership Team: This workshop introduced the participating teachers to the main objectives of the project.
4. Coaching sessions for teachers in the use of IR: This amounted to six school visits, including the introduction workshop above. The timing of these sessions varied between schools, depending on their schedule. The sessions aimed to provide support with IR, as well as further support related to curriculum planning and assessment in the context of IR usage.
5. Online support (through AfA's 'The Bubble'[10] online platform) for schools containing materials on collaborative action research, literacy and English, and identifying and supporting speech, language and communication needs. During the initial start-up phase a series of technical (rather than pedagogical) challenges were faced. As a result, instead of using 'The Bubble', coaches shared observations, challenges, reflections and ideas through monthly in-house conference calls, supported by a comprehensive suite of Office 365 collaborative tools (Skype for Business, Sharepoint, OneNote, etc.) to inform their supportive dialogue with schools.

Evaluation methodology

This section outlines the specific questions this report seeks to address and the methods of doing so. The procedure and changes to the procedure are described in detail.

Research questions

This study investigated the effects of IR on reading outcomes as well as the process of implementing IR in participating classes. The study addressed several research questions, as follows.

The impact evaluation component of this study was designed to investigate the following research hypothesis:

- IR will have a positive effect on the outcomes in reading comprehension of pupils who are randomly assigned to the intervention condition compared to pupils in the control condition.

The IPE was designed to address the following questions:

- Was the intervention implemented with fidelity in the intervention classrooms?
- What factors and initial conditions appear to explain variation in fidelity of implementation?
- What appear to be the necessary conditions for success of the intervention?
- What were the barriers to delivery?

10 http://res.afa3as.org.uk/BUBBLE/TheBubble/index.html

Table 1 Randomisation result: class allocations by year group

	Intervention	Control	Total
Year 3	20	20	40
Year 4	20	20	40
Total	40	40	80

Note: After randomisation, one school had a combined Year 3 and Year 4 class which, for the purposes of analysis later on, was treated as two separate classes.

Study design

This study consisted of a two-arm cluster[11] RCT with an IPE component. The unit of randomisation was Year 3 and Year 4 classes within schools. All Year 3 and Year 4 classes in participating schools were randomly allocated to either the intervention or the control condition and each class had a 50 per cent chance of being selected to the intervention condition. Classes in the intervention condition received the IR package outlined above, while classes in the control condition continued with 'business as usual' and did not have access to IR or any training in its use.

Randomisation

Randomisation took place on the 6 July 2018 and included all classes for which the schools had signed a participation agreement (a memorandum of understanding). In each school, the Year 3 and Year 4 groups had two classes respectively (i.e. two classes per year group in each school). The process was to create random numbers for each unit; rank all classes according to the random number and create a ranking variable; and then use the school, year and ranking variable to sort data by school. The first ranked class in each year group was allocated to the intervention and the remaining class was allocated to control. One school (i.e. four classes) did not have a participation agreement at that point. To avoid randomising at a later date, the school was included in the randomisation process but the allocation result was not revealed to AfA until after the school had submitted the paperwork. The randomisation resulted in 20 Year 3 classes and 20 Year 4 classes being allocated to the intervention condition, and 20 Year 3 classes and 20 Year 4 classes being allocated to the control condition (Table 1).

Participants and trial attrition

This section describes the study sample and dropout during the study.

Recruitment

Initially, the aim was to recruit 20 schools and 80 classes, with 40 classes in each trial condition. Based on an assumption of 20 pupils per class,[12] the evaluation team expected at least 1,600 pupils to be included in the study. Only Year 3 and Year 4 teachers in participating schools were included in the trial.

11 Pupils are part of (nested in) school classes, which are in turn nested in schools. The classes ("clusters") containing multiple pupils are randomly assigned to treatment and control group.

12 In 2018, over 33 per cent of teachers in primary schools now teach classes with more than 30 pupils. Teacher Tapp. 2018). 'What Teachers Tapped This Week.' As of 26 August 2019: https://teachertapp.co.uk/what-teachers-tapped-this-week-62-3rd-december-2018/

Schools were recruited by AfA based on their capacity to recruit and provide training. For schools to be included in the trial, the following inclusion criteria were considered:

- Location: Schools were selected based on the AfA's capacity to provide training; IR coaches were broadly located across south-west England.
- IT resources: To participate, schools needed to be able to provide pupils with 1:1 access to IR-enabled devices, such as desktop computers, laptops or tablets.

The following exclusion criteria were also considered:

- Involvement with AfA: Schools could not participate in the trial if they were participating in other programmes by the AfA delivery team at the time of recruitment.
- Involvement in other reading or literacy trials: Schools could not participate in this study if they were currently involved in other reading or literacy trials or interventions.

In total, 20 schools fulfilled the above requirements and were recruited in June 2018 to participate in the trial. These 20 schools were randomly allocated to the intervention and control groups following the procedure outlined above. One school had a mixed Year 3 and Year 4 class which, for the purposes of the analysis in this report, was considered as two separate classes (with the appropriate reading measure administered).

School attrition

Seven out of the 20 schools initially recruited dropped out of the trial. This occurred as follows: four schools between July and December 2018; one school in January 2019; one school in April 2019; and one school in May 2019. The first six schools dropped out from the trial before any baseline data could be collected, while the seventh was not responsive to the invitation for endline data collection. Although not all schools provided reasons for dropout, those which did mentioned:

- Project too time-intensive or teachers generally did not have the time to implement
- The 'robotic nature of the voice' used by IR was reported by teachers as potentially problematic
- Issues with IT access and log-ins (to both IR and the reading assessment)
- The school underestimated the level of equipment and software needed (e.g. Windows 10).

Therefore, from randomisation to analysis stage, there was 35 per cent attrition at the school level and 33 per cent attrition at the class level (given the fact that the class numbers increased by two when the mixed Year 3 and Year 4 class was separated into the two year groups, as above).

Outcome measure

The outcome of interest for this study was reading comprehension. It was assessed using the Progress in Reading Assessment (PiRA)[13] test, developed by Hodder Education. This is a termly test which offers full coverage of the UK curriculum.[14] It has been shown to have high internal validity (or test reliability)[15] and has a high correlation with external measures of

13 https://www.risingstars-uk.com/series/assessment/rising-stars-pira-tests

14 https://www.gov.uk/government/publications/national-curriculum-in-england-framework-for-key-stages-1-to-4

15 Cronbach's alpha of above 0.9.

Table 2 Endline outcome collection by year group

	Intervention	Control	Total
Year 3	13	13	26
Year 4	14	14	28
Total	27	27	54

attainment.[16] The test provides standardised results and age-related scores that are benchmarked across national averages. Participating pupils received test versions corresponding to their year group. Tests were supposed to be delivered and marked independently by Rising Stars Assessment; however, due to technical issues with the online testing platform at baseline, AfA collected data using the equivalent paper version of the PiRA assessment. Employees of AfA also marked the tests and entered the data, as opposed to this being done by Rising Stars Assessment (the developers of the PiRA instrument). The class-level aggregated and anonymised data was then sent to the evaluation team for analysis, securing an independent evaluation.

Timing of reading assessments

Initially, reading comprehension was to be assessed on three separate occasions:

First, reading comprehension was measured at baseline, initially scheduled for September and October 2018 (before IR implementation in intervention classes). AfA liaised with Rising Stars Assessment for the carrying out of the baseline data collection. Technical issues with the testing platform led to some schools only completing the baseline assessment once the intervention had started, which led to the abovementioned use of paper versions of the test. There was also insufficient information as to the exact timing of the baseline data collection and implementation start in all schools. As a result, the collected baseline data could not reliably be used as a measure of reading prior to any IR implementation and the measure was therefore removed from the analysis.

A second data collection point was scheduled for the midline, initially December 2018. Due to technical issues with the testing platform which cause delays with the baseline testing, this data collection point was compromised. AfA staff put in considerable effort to address this, by using the paper version of PiRA instead of the online version, and hand-marking all tests. However, this midline data collection happened too close to the baseline data collection point in a sufficient number of schools that it could not reliably be considered a true midline in relation to implementation. Therefore, the decision was made by the evaluation team to not use these data in the final analysis.

The third and final data collection point was at endline in May and June 2019. This testing took place in all schools still participating in the trial at that point. A total of 1,220 pupils across 54 classes completed the reading assessment in May 2019 (see Table 2 for an explanation of class numbers in sample). The evaluation team were not provided with all data on the number of pupils per class (data for one school was missing); it has therefore not been possible to calculate an overall pupil response and attrition rate. However, from

16 McCarty, C. & K. Ruttle. 2018. *Progress in Reading Assessment. Manual (Stage 2). Second Edition*. Hodder Education.

Table 3 Class reading comprehension level (PiRA) at baseline

	Intervention	Control
Standardised reading score at baseline	97.22 (SD=6.96)	97.90 (SD=6.61)
N (classes) **N total = 58 classes**	29	29

available data, an average of 86.69 per cent of pupils per class completed endline testing, with rates varying from 63.33 per cent up to 100 per cent. With the data analysis to be undertaken at class level, the pupil-level attrition does not affect the analytical approach outlined in what follows. However, this attrition may have implications for the internal validity of the trial, as the characteristics of pupils withdrawing from the endline outcome testing within each participating class were not planned to collect and are, therefore, unknown.

Reading levels at baseline

The baseline reading assessment (supported by AfA) took place between October 2018 and February 2019. A total of 14 schools with 58 classes took part in this assessment. This was initially conceptualised as constituting a baseline measure for the purposes of the final analysis, but was not used as such due to delayed baseline testing. However, the results of that assessment are included here for completeness. Given the random assignment to the intervention and control conditions, any differences in the two groups are due to chance and therefore statistical tests of differences between means are not appropriate. Instead guidelines that suggest only simple comparisons should be undertaken were followed. The data indicates that there were no differences in reading comprehension between classes in the control group and those in the intervention group.

Trial design update

The data collection difficulties led to a change in the trial design, as outlined in Table 4. Changes include a smaller number of participants, PiRA being assessed by AfA instead of Rising Stars Assessment and the use of endline data only for the quantitative analysis.

The trial retained its two-arm cluster-randomised stratified nature; however, the loss of baseline and midline data had implications for specific hypotheses that could be tested and for the statistical analysis, as follows.

Statistical analysis

The analysis of outcomes was undertaken on an intention-to-treat (ITT)[17] basis. This method compares outcome means for the intervention and comparison groups, and participants are analysed according to their randomised group allocation, regardless of whether or not they adhered to this group. The ITT approach is inherently conservative as it captures the averaged effect of *offering* the intervention, regardless of whether participants comply with the assignment.

The unit of analysis was the classroom. To explicitly account for the clustering of classes into schools, multi-level regression models

17 Gupta, S. K. (2011). Intention-to-treat concept: a review. *Perspectives in clinical research*, 2(3), 109.

Table 4 Original and updated evaluation design

	Original design	Updated design
Trial design	Two-arm, stratified, cluster RCT with allocation at class level	NA – same design
Pupil age range and key stage	7–9 years old; Year 3 and Year 4; Lower Key Stage 2	NA – same age range
Number of schools	Recruited: 20 schools (80 classes) At baseline: 14 schools	Tested: 13 schools (54 classes)
Number of pupils	Estimate: 1,600 (assuming 20 pupils/class) Actual at baseline: 1,364 (completed baseline testing)	1,220 (completed reading test at endline)
Primary outcome	Reading	NA – same outcome
Primary outcome measure	PiRA test by Rising Stars Assessment	PiRA test by Rising Stars Assessment, but collected and marked by AfA
Measurement timing	Baseline; Midline; Endline	Endline only

Note: Please refer to Annex for a more detailed explanation on number of classes presented in this table.

are usually employed. However, given the small number of schools in this trial, the evaluation team has accounted for the school-level clustering via a school fixed effects specification. Linear fixed effects regression models[18] make no assumptions about *what* internal school aspect may affect pupils' outcomes, essentially differentiating out the entire school effect on outcomes, regardless of what might be driving it. All standard errors were clustered at the school level.

The original analysis plan foresaw the use of these school fixed effects linear regression models to ascertain the impact of IR on:

- Pupils' reading levels, using baseline, midline, and endline data.
- Pupils' reading progress, i.e. the change in reading levels from baseline to midline, and from baseline to endline respectively.

Due to the trial changes, the final statistical analysis used a school fixed effects linear regression model to ascertain the impact of IR on:

- Pupils' reading level at endline.

Although this means that the evaluation cannot ascertain the impact on pupils' progress in reading, the overall impact of IR on reading levels can still be estimated. The results are also presented in terms of an effect size, which allows for comparisons across other similar intervention and RCTs. The exact model specification and the calculation method for the effect size are included in Annex A.

18 Clarke, P., C. Crawford, F. Steele & A. Vignoles. 2015.' Revisiting fixed-and random-effects models: some considerations for policy-relevant education research.' *Education Economics*, 23(3),): 259–77.

Table 5 Survey response rates

	Head teachers		Intervention teachers		Control teachers	
	No. sent to	*No. responses*	*No. sent to*	*No. responses*	*No. sent to*	*No. responses*
Baseline	19	7	39	8	42	3
Endline	14	4	26	9	26	11

Note: Surveys were distributed by the delivery team to all schools actively participating in the trial at the time the surveys were administered. At baseline, data collection surveys were sent to 19 schools. One school dropped out at endline stage, resulting in a total sample of 13 schools for the endline survey. Please note that timing for the survey data collection did not coincide with the data collection for outcome data.

Implementation and process evaluation methods

Accompanying the impact evaluation, implementation and process evaluation data was collected to shed light on how IR was implemented in intervention classes. The IPE looked at *implementation fidelity* to identify deviations to the intervention (i.e. to what extent was the intervention conducted as planned), the *necessary conditions* to successfully implement IR in the classroom and *barriers* that teachers and coaches encountered. IPE data was collected through online surveys and semi-structured interviews with staff in two case study schools.

Surveys were sent to head teachers and control and intervention teachers in November 2018 (baseline) and towards the end of the intervention in June 2019 (endline). The baseline survey asked all respondents (control and intervention teachers and head teachers) about their prior experience using technology. Intervention teachers, as well as all head teachers, were also asked about their experience with the training and level of preparedness to implement the intervention. The endline survey asked intervention teachers and head teachers about their experience implementing IR, the coaching and other support received, and about their perceived impact of the intervention on pupils. Control teachers were asked about their practices using technology in the classroom in the past year. The online surveys were distributed by the delivery team on behalf of the evaluation team. The head teacher survey had a small number of responses at endline, and therefore the perspectives of these respondents should not be taken to be representative of the whole participating sample. Response rates for the intervention teachers and the control teachers ranged between 7 per cent (control teachers at baseline) and 37 per cent (head teachers at baseline). Table 5 details the number of responses for each survey.

Case-studies with two schools were initially planned, with a random selection procedure from among volunteering participating schools. However, only two schools volunteered to take part in the case studies.[19] Semi-structured telephone interviews with teachers and head teachers from these two case study schools were conducted to gain further insights on their experience using the IR tool. Four interviews were conducted in total, two with head teachers and two with teachers (one from each of the case study schools).

Finally, coaching records (coach-provided individual records of meetings) were collected

19 This may introduce bias in the IPE as schools volunteering may have had extreme (either positive or negative) IR experiences.

during the course of the trial. These coaching records provided information for the purposes of monitoring programme participation, in particular information about teachers' engagement with training and coaching sessions, which were in turn collated by AfA Project Lead to inform AfA Reporting.

Ethics and data protection

To protect the privacy and rights of study participants, ethical clearance was sought from the RAND US Human Subjects Protection Committee (HSPC) and all evaluation procedures adhered to data protection laws.

The study was approved by HSPC (IRB Number IRB00000051). This required obtaining opt-in consent from teachers and head teachers in selected schools to participate in the IPE interviews, and opt-out consent from the additional testing for the impact evaluation component from pupils' parents or legal guardians. With the intervention and the testing taking place within the school day, schools act *in loco parentis* and act as gatekeepers to all pupils.

Schools and parents were provided with fair processing privacy notices that explain the use, storage, and secure handling of the data for the entire study. RAND Europe adopts good industry practices regarding the protection of personal data as part of its obligations as a data controller under the Data Protection Act 1998 and takes appropriate technical and organisational measures conformant with ISO 27001 to protect personal data.

RAND Europe did not collect any individual-level pupil data. All pupil-level data was collected by AfA. AfA were responsible for aggregating the data at classroom level and ensuring the evaluation team did not receive any identifiable data (pupils, teachers or school names).

Results: Impact evaluation

Impact of IR on reading outcomes

The school fixed effects model was specified using dummy variables for all schools but one (the reference category was chosen at random). The trial result is provided by the coefficient of a separate variable which identifies allocation to intervention conditions (coded 1 if the class was randomly allocated to the intervention condition, and 0 if it was allocated to the control condition). There were no other predictors in the model.

Given the technical timing issues with the baseline reading assessment, and the decision to not use that data, the model did not include a baseline measure.

The result of this model points to a small, negative, but statistically non-significant difference in the reading levels between classes allocated to the IR intervention and control group respectively.

Table 6 Class reading comprehension level (PiRA) at endline

	Intervention: Unadjusted mean (standard deviation)	Control: Unadjusted mean (standard deviation)	Unstandardised coefficient for allocation variable (confidence interval)	Significance (p-value)	Partial eta squared (η2) effect size
Outcome: Standardised reading score at endline	M=100.54 (SD=6.03)	M=101.68 (SD=3.86)	-1.132 (-4.175 1.910)	0.456	-0.014
N (classes) **N total = 54 classes**	27	27			

This suggests that IR classes performed as well as non-IR classes in the final reading attainment assessment. The school-level attrition is likely to have affected the precision of this estimate. The reasons for schools withdrawing from the trial are known (detailed in the *Participants and trial attrition* section); however, in the absence of specific information as to the characteristics of the pupils in the schools having dropped out from the trial, the direction of any emerging bias cannot be estimated. Overall, the impact of the high level of class attrition is to reduce the internal validity of the study, and in conjunction with the implementation and process evaluation results below highlights the importance of interpreting this result in light of both attrition and implementation experiences.

Results: Implementation and process evaluation

Deviations from initial implementation plan

Data collected as part of the IPE provides insights into how IR was implemented in practice. In particular, it provides details on the fidelity of implementation, as wells as enablers and barriers teachers and schools experienced while using it. As such, IPE findings can provide plausible explanations for the null outcome evaluation results.

Intervention monitoring information revealed divergences from the original intervention design, as follows:

- Intervention duration: IR was intended to run for eight months per school. However, crashes of the PiRA platform caused significant delays. The actual starting dates in schools are not known and implementation length likely differed between schools by up to four months.
- Coaching: While AfA reports that all schools had six visits by coaches, the timing of these differed (e.g. some coaches combined the content of multiple visits into a longer one to accommodate school schedules). Additionally, AfA reports that some coaches were involved in a larger capacity (timewise) than planned, largely due to IT and equipment problems. This is a deviation from the original intervention design. This was confirmed by at least one interview.
- Online support: Online support for teaching with the IR was supposed to be available via AfA's platform 'The Bubble'. This was not made available to participating teachers. During the initial start-up phase a series of technical (rather than pedagogical) challenges were faced. As a result, instead of using 'The Bubble', coaches shared observations, challenges, reflections and ideas through monthly in-house conference calls, supported by a comprehensive suite of Office 365 collaborative tools (Skype for Business, Sharepoint, OneNote, etc.) to inform their supportive dialogue with schools. Many of the resources shared with schools in this manner were developed by the delivery team's Project Lead, who also spent considerable time liaising with Microsoft Tech teams to overcome Office 365 deployment, sign-in and other issues, communicating directly to coaches and school teachers, leaders and technical support, producing instructional videos as well as notes.

Initial expectations and training

Teachers' initial expectations about IR, as reported in the baseline survey, were largely positive. They also exhibited an overwhelmingly positive attitude towards the potential of using technology to improve educational outcomes in schools. Furthermore, all baseline survey respondents claimed familiarity with using technology in the classroom, as well as with the use of Microsoft Office software. In the interviews conducted after the intervention, one head teacher noted that their motivation to join the project derived from the fact that IR was a

new and unfamiliar technology and staff wanted to explore the opportunities it presented.

However, the baseline surveys for intervention teachers revealed a degree of confusion among teachers concerning both the purpose of IR and the requirements for them to use the tool. Only some indicated that they were 'clear' or 'very clear' about the requirements. There was also a mixed response concerning implementation, with half of the baseline teacher survey respondents reporting that they were not very clear as to how IR should be implemented. This is as expected, as the survey was intentionally sent out before teachers met coaches for the first time.

In terms of the training and coaching, survey responses revealed that there were low levels of uptake and mixed views on the usefulness of the sessions and how well these prepared teachers to use IR. At endline, only three out of nine intervention teachers reported that they had attended the introduction workshop. The delivery team monitoring data (collated through the coaching records) suggests, however, that 26 teachers received face-to-face training and follow-up by coaches. While two out of these three respondents agreed or strongly agreed that they were able to ask questions specific to their school during the session, they expressed less positive views in terms of the clarity of information and usefulness of the training (see Annex B, Table A for details). Moreover, 50 per cent of surveyed intervention teachers (four out of eight, one respondent skipped the question) reported feeling 'somewhat' prepared to implement IR. Only one respondent reported that the training had prepared them 'very well', while the remaining three reported feeling that the training prepared them 'not very well' (two) or 'not at all' (one). In addition, as part of the intervention, teachers had coaching sessions available to them. In the endline survey, only one out of the nine intervention teacher survey respondents said that they had attended online coaching sessions. Survey responses on the uptake of face-to-face coaching were inconclusive. However, a teacher and head teacher from the same school reported valuing the support the IR coach provided them, but expressed a preference for more face-to-face support over online coaching (T1, HT1).[20]

Intervention teachers were asked to rank their overall preparedness to implement IR in the classroom on a 1-10 scale (where 1 indicates 'completely unprepared' and 10 'fully prepared'). The mean score of the nine surveyed teachers was 4.22 and the median was 3 (scores ranging from 2 to 10). More than half (seven out of nine) of the respondents ranked their level of preparedness with a 4 or below, showing that most respondents felt unprepared rather than prepared to start using IR. In line with these findings, one of the interviewed teachers commented that 'the training didn't really teach us to deliver the intervention; it just showed us how to use Immersive Reader' (T1). These findings suggest that improvements can be made in the content of the training and in the ongoing support provided to teachers to use the IR intervention. The delivery team report (summarised in Box 1) identifies a lack of engagement from some participating teachers with the professional development provision generally as a barrier faced by coaches when delivering the training. Head teachers expressed similar, though slightly more positive, views on their school's perceived level of preparedness to introduce IR following the introductory workshop. Most head teachers rated their school's preparedness with a 5 or below (five out of eight head teachers), suggesting higher levels of feeling unprepared. However, three head teachers rated their school's level of preparedness more positively (two head

20 Throughout this section we refer to the interviews as follows: case study school A: HT1, T1; case study school B: HT2, T2. For instance, HT1 refers to the head teacher in School A.

teachers rated it as a 7, and one a 9). By the end of the intervention, these mixed views were corroborated by the interviewed head teachers. While one interviewed teacher commented on the useful content and structure of the training sessions in preparing school staff to use the tool (T1), the head teacher of the same school reported that their teachers mainly learned how to use the tool through 'playing with it themselves' (HT1). In fact, both interviewed head teachers suggested that more time and earlier training would have been beneficial, as they were mostly unclear about the technical implications of implementing IR (HT1, HT2).

Initial implementation and daily use of IR

Survey responses and interviews revealed that teachers experienced difficulties while using IR due to a lack of resources and IT infrastructure at their schools. These views were corroborated by head teacher reports. Additionally, teachers also reported that IR was difficult to integrate in the classroom. This was partly due to resource and technical difficulties, but also due to the IR software being difficult to use both by teachers and pupils. Finally, IPE data suggests that using IR is time-consuming and requires technologically savvy individuals.

Endline survey teacher respondents felt that they did not have the resources required to use the IR tool effectively with their classes. All nine surveyed teachers at endline reported that they did not have enough computers or tablets to use IR. Head teachers also recognised this as a barrier; three out of four head teachers who answered the endline survey agreed that their staff felt that they did not have access to enough tablets or computers for pupils, and reported that upgrades to their school's IT infrastructure were necessary in order to implement IR. This is also reflected in findings by AfA (see Box 1).

Considerable technical difficulties were experienced by teachers and they felt IR was difficult to integrate into the classroom. One teacher reported poor internet connection in the classrooms during the interview (T2). Other problems concerned the software itself; interviewees in both schools highlighted that technical issues relating to the use of IR on iPads were key barriers to implementing the intervention successfully (HT1, HT2, T1). One teacher also highlighted a number of difficulties in getting IR to work within their school's network as the school did not subscribe to the latest version of Office365 (T1). Interviewees noted that the IR would not work effectively or consistently with different equipment (e.g. tablets compared to laptops) (HT1, T1, HT2, T2).

There was almost universal agreement by teachers responding to the endline survey that it was difficult to integrate IR into their daily teaching practices. Survey responses concerning how pupils interacted with IR offer further explanation as to why this was the case. The majority of teachers (six out of nine) voiced agreement with the statement that it was challenging to get pupils to use IR and that pupils got easily distracted by the features of the IR app. One teacher reported pupils needed to gain confidence in using IR, and only once this had been achieved could they use it independently.

Teachers reported that the implementation of IR negatively affected their workload. In the endline survey, six out of nine intervention teachers strongly agreed or agreed that implementing the IR tool required them to work overtime. Furthermore, following the implementation of IR, teacher views on whether technology could save teaching time changed. While in the baseline survey the majority of teacher respondents (five out of nine) disagreed that using technology for reading takes more time than traditional methods, following the trial teachers were almost in unanimous agreement that technology-aided reading teaching took more time than traditional methods, including IR specifically (seven out of nine intervention teachers agreed with this). In the interviews, one

teacher described the preparation required to use IR as 'quite onerous' and noted that it would have negatively affected their workload had they used it regularly (T2). Another teacher described the process of getting the IR tool to operate within the school system as 'very complex' (T1). The need for additional staff time to implement IR was also noted by one interviewed head teacher (HT1). Finally, the head teacher and teacher from the same school noted that having in-house IT support was paramount for resolving IT infrastructural issues and identified this as a necessary feature to facilitate the use of IR (HT1, T1).

Perceived impact

Teachers were largely sceptical of IR's impact on improving pupils' reading ability, but views varied. Despite a generally positive attitude towards the use of technology in reading instruction, teachers were mostly negative in their assessment of the impact of the IR tool on improving both pupil's reading ability and engagement in reading; only one teacher out of nine in the endline survey agreed that the use of the IR tool improves pupil's ability to read and comprehend text (six out of nine disagreed and two respondents did not know).

Endline survey responses indicated that teachers were generally ambivalent about whether IR would be effective for group instruction, pupils with special educational needs (SEN) or pupils that had English as an additional language (EAL). Similar mixed views were voiced in the interviews. One interviewed teacher noted that IR went against their school's approach to reading, which they described as 'geared towards reading with expression and fluency' (T2). As a result, they were wary of using IR with their whole class because, in their opinion, the programme's voice was not expressive enough when reading (the intonation came across as 'robotic') and therefore affects how the pupil understands the text (T2).

Still, this teacher suggested that IR could be beneficial for EAL pupils or for children with low reading ability (T2). The teacher and head teacher from another school (T1 and HT1) also saw some potential in IR (despite not having seen impacts on their pupils). The head teacher noted that they 'could see how [features of IR] could benefit children with dyslexia or visual difficulties, […] children with focus problems, […] EAL children and children with language processing difficulties' (HT1). This head teacher commented that they might continue using IR at their school but only with individual children if it is appropriate for their needs.

There was no clear consensus between teachers as to which learning setting the IR tool would be most effective in, the features of the tool they used most frequently or the age group for which using IR would be most effective. In the endline survey, 'small group instruction' was cited most often by teachers (four out of nine) as 'somewhat likely' to be a setting where IR would be effective, while using IR for independent work at home was seen as the least appropriate environment for IR. In the endline survey, teachers' responses on which features of the IR they used most frequently were mixed. Only one feature – identifying parts of speech – was used frequently by more than half of teachers (five out of nine teachers reported using it frequently or very frequently). The survey data was also inconclusive as to with which age group IR was deemed to be most effective by teachers.

Delivery team (AfA) findings

Throughout the trial, the delivery team collected views from teachers, head teachers, and coaches on the IR intervention. Box 1 lists the findings that consistently emerged in both the evaluation IPE and in AfA's monitoring data.

Box 1 Results from AfA that match evaluation findings

- Findings from both the AfA report and IPE data indicate that school staff had particular preference for **direct, face-to-face training and support**. Additionally, both sources identify that improvements could be made in terms of the **relevance of the training** and **quantity and type of ongoing support provided to schools** (e.g. support for using IR in the classroom and more general IT support).
- In line with the above, both sources also highlight the fact that not all UK schools have the necessary **quality and quantity of IT resources, infrastructure, and staff with IT expertise** to support the delivery of the intervention. This is an important consideration should there be intentions to roll out the IR tool more widely in UK schools.
- Both reports identified that trial participants had difficulties with the IR tool when using it with different **platforms and devices.** The inconsistency in how the software work with different devices impeded the smooth delivery of the intervention.
- Additionally, findings from both the AfA report and IPE show that school staff faced challenges in accessing **Microsoft Office 365 as some schools had to create online accounts and logins for staff and pupils to access the IR tool**. This suggests that despite the fact that one of the requirements for being recruited to participate in the trial was having access to Office365, it seems that schools had not fully understood the requirement and confused Office365 with Microsoft Office suite (e.g. having access to Microsoft Word).
- In terms of the IR tool itself, both reports identified that school staff found the **computerised voice** of the IR tool to be a limiting factor for using it with their pupils. They explained that the voice lacked the **expression and intonation** required for effective reading instruction.
- Furthermore, findings from the AfA report and IPE indicate that students would have benefited from further **guidance and support** to use the tool throughout the intervention.
- Findings from the AfA report indicate that **EAL pupils** were supported by the translating function of the IR tool. While IPE data yields mixed results in terms of the opinions of teachers and head teachers in this regard, there were positive expectations on the potential benefits of using the IR tool with **low reading ability** and **EAL** children.

Source: Achievement for All. 2019. Immersive Reader Delivery Team Report.

Conclusion

The main outcome analysis of this two-arm RCT does not show a statistically significant effect of IR on reading outcomes. This suggests that classes that received IR perform as well as classes in the 'business-as-usual' condition. The trial was affected by school-level attrition, with 35 per cent of schools initially recruited dropping out, and also by a change in the outcome measure collection schedule, which resulted in only one hypothesis concerning the effects of IR being tested.

The IPE provides several insights into the day-to-day implementation and usage of IR. The main finding of this process component relates to the inconsistencies in the day-to-day implementation of IR in terms of length of implementation, coaching involvement and

the online support for teachers participating in the intervention. This suggests that the intervention was not implemented as intended, nor was there uniformity in participating schools' reported practices. This added a further challenge to the internal validity of the study, and any future evaluations of similar interventions, including IR, would benefit from the careful monitoring of implementation processes and fidelity.

Teachers' initial expectations about IR were largely positive. They exhibited an overwhelmingly positive attitude towards the potential of using technology to improve educational outcomes in schools. However, there was some confusion among teachers concerning both the purpose of IR and the requirements for them to use the tool. They generally considered themselves rather unprepared.

Teachers experienced difficulties while using IR due to a lack of resources and IT infrastructure (such as operating systems, computers, tablets, etc.) at their schools and thought it was difficult to implement in the classroom. This is in contract to findings by the delivery team in relation to schools' preparedness to implement the IR package (as outlined in Box 1). These suggested that the initial assessments that head teachers provided about their schools' readiness in terms of IT systems did not initially raise any concerns that the available IT provision might be insufficient for the successful implementation of IR. The delivery team noted in their report that a potential solution to this issue is to implement a technical visit to participating schools prior to implementation; the visit would more accurately assess the availability and readiness of the technical resources required to support a successful implementation. Teachers were also largely sceptical of IR's impact on improving pupils' reading ability. Despite a generally positive attitude towards the use of technology in reading instruction, teachers were mostly negative in their assessment of the impact of the IR tool on improving both pupil's reading ability and engagement in reading. Some mentioned, however, that IR might be useful for SEN or EAL children.

Taken as a whole, the IPE results suggest that greater coherence in designing and implementing the intervention would be beneficial for a re-run of this trial or further evaluation of IR and any other similar interventions. Any future implementation and testing of IR should consider several aspects. First, issues of implementation clarity early on, so that schools and teachers are aware of all the implications of participation. A specific challenge will be to assess the actual IT resources of schools. Secondly, implementation consistency across participating schools, classes and teachers, to include a measure of implementation fidelity, to ensure that any effectiveness study is concerned with one consistent intervention that adheres to the design of the intervention. And lastly, support to schools and teachers to deal with technical difficulties that may impede the implementation of the intervention. With implementation issues addressed, future trials can reassess the impact IR may have on pupils' reading outcomes.

Annex A

Note on groups

One school had mixed year classes. Data was analysed as provided by AfA:

'Year 3 = 1 x Controlled Class and 1 x Treatment [intervention] Class. Year 4 = 1 x Controlled Class and 1 x Treatment [intervention] Class = 4 classes in total. [But] each class has mixed Year 3 and 4 students so there are two sets of age-specific test results for [each] class. This equates to 2 sets of data per class – or 8 sets of data for the school.' (Personal Correspondence). The number of classes in text is provided as analysed.

Statistical analysis

The general equation for the school- fixed effects model is given below:

$$y_{ij} = \alpha + Z_i b_j + \delta DF_j + + M_j + u_{ij} \qquad i = 1..N, j = 1..M, \qquad (1)$$

Where y_{ij} denotes the outcome of classroom i in school j. Normally an X_{ij} $1 \times k$ matrix of classroom characteristics, including baseline attainment data, would be included in the model. Since no reliable baseline data is available, this has been excluded from the model. DF_j is a dummy variable denoting the intervention effect and M_j are n-1 dummy variables for school fixed effects. Finally, u_{ij} is the school-level error term. In equation (1), α denotes the overall model intercept. u_j are referred as school-level residuals ($u_{ij} \sim i.i.d\ N(0, \sigma^2)$).

The coefficient δ associated with the DF_j dummy will constitute the main result of the trial. To obtain the main result of the trial, Equation (1) above will be estimated for all pupils in all classes.

Effect size calculation

The effect size is reported as a partial eta squared. The partial eta squared is a measure in which the effects of other independent variables in the model and interactions are partialled out.[21]

The calculation of the effect size (partial eta squared) used the following formula:

$\eta 2$ = Sum of Squares (between groups) / [Sum of Squares (between groups) + Sum of squares (within groups)]

Where Sum of Squares is the sum of the square of variation for the outcome variable.

21 Richardson, J.T. 2011. 'Eta squared and partial eta squared as measures of effect size in educational research.' Educational Research Review, 6(2),): 135–47.

Annex B

Table A Detailed endline survey responses on initial training attendance (intervention teachers)

Have you attended the introduction workshop or the theory of change workshop (ToC) of the Immersive Reader (IR) training sessions?						
Yes	3					
No	6					
	Answered: 9					
	Skipped: 0					
***Sub-question:* Based on your attendance to the introduction workshop, to what extent do you agree with the following statements?**						
	Strongly agree	**Agree**	**Disagree**	**Strongly disagree**	**Don't know**	**Response total**
1. The information provided was unclear.	33.3% (1)	33.3% (1)	33.3% (1)	0.0% (0)	0.0% (0)	3
2. The information provided was useful.	0.0% (0)	0.0% (0)	66.7% (2)	0.0% (0)	33.3% (1)	3
3. The information provided was detailed enough.	33.3% (1)	0.0% (0)	33.3% (1)	33.3% (1)	0.0% (0)	3
4. It was possible to ask questions specific to our school.	33.3% (1)	33.3% (1)	0.0% (0)	0.0% (0)	33.3% (1)	3
5. The Immersive Reader team provided useful answers to my questions.	0.0% (0)	0.0% (0)	66.7% (2)	0.0% (0)	33.3% (1)	3
6. The Immersive Reader team was knowledgeable about the practicalities of using IR.	33.3% (1)	0.0% (0)	66.7% (2)	0.0% (0)	0.0% (0)	3
					Answered	*3*
					Skipped	*6*