Research Article

Evaluation of the impact of Accelerated Reader on English Reading Performance and Behaviours in Chinese Primary Schools – A Pilot Study

Fujia Yang* [ID] and Beng Huat See

School of Education, University of Birmingham, Birmingham B15 2SA, UK

https://www.engineegroup.us

Check for updates

Abstract

This paper reports on the pilot results of the first independent evaluation of Accelerated Reader (AR), an online reading programme, in China. Despite its adoption in over 800 Chinese schools and robust evaluation elsewhere, AR has not been independently assessed in China. The sample included 528 Year 5 and 6 pupils from two public schools in China. The pilot was a cluster randomised control trial, where four classes (195 pupils) were randomised to receive the AR intervention, while seven (333 pupils) followed business-as-usual instruction. The intervention lasted 12-13 weeks, with one session per week. Impact evaluation showed positive effects on English reading outcome (effect size [ES] = +0.27), overall reading habits (ES = +0.14) and attitudes (ES = +0.15), though regression models suggested these may reflect pre-existing differences.

Compliance analysis showed that pupils who complied made greater progress than non-compliers (ES = +0.56), highlighting the importance of session completion. Process evaluation reflected large variation in implementation fidelity, driven by teacher experience, classroom management, and technical support. Key challenges included pupils' limited English proficiency, digital skills, large class sizes, and mismatched book access.

Path analysis indicated a small positive indirect impact of reading behaviours on performance, mainly through reading attitudes.

Introduction

China's political reforms and open-door policy in the last 46 years have led to a number of important education reforms. As part of the modernisation strategy, English emerged as a core subject along with maths and Chinese. Since 2003, it has been introduced as a compulsory subject from Year Three (ages 8-9) in primary schools, although many schools have gradually introduced it earlier, starting from Year One [1]. In recent years, over 107 million primary school pupils across the country have engaged in learning English as a foreign language [2]. To facilitate this large-scale educational effort, the MOE implemented a structured approach to English education, including the development of curriculum standards and standardised learning materials. However, despite these nationwide efforts, this "one-size-fits-all" approach in choosing reading resources may fail to address the diverse proficiency levels present within classrooms [3,4]. While standardised resources aim to streamline the teaching process, they often overlook the significant variations in individual pupils' learning needs.

Moreover, existing research has highlighted the vital connection between proper reading resource availability and children's academic performance. Mullis, et al.'s [5] analysis of the Progress in International Reading Literacy Study (PIRLS) 2021[6] found that pupils attending well-resourced schools tended to excel in reading assessments. This highlights the urgent need for the provision of adequate reading resources to support learners of diverse ability.

Yet, providing adequate resources alone is not enough—children's reading behaviours, including their habits and attitudes, also play a crucial role in their reading performance.

Longitudinal studies have shown that reading habits become a stable predictor of reading comprehension from upper primary years onward [7,8], supporting the idea that consistent engagement with texts fosters long-term gains. In contrast, reading attitudes appear to decline with age and show weaker or inconsistent links to achievement [9-11]. This suggests that interventions focused solely on improving attitudes may be insufficient, particularly in later primary years, unless they are accompanied by efforts to cultivate consistent reading habits.

In response to these challenges, numerous reading interventions, such as Reading Recovery, Reciprocal Reading, Guided Reading, and Accelerated Reader, have been developed to support pupils' reading performance. Among the reading interventions that have been rigorously evaluated, Accelerated Reader (AR) appears to show promise with a number of studies reporting positive results [12-16].

While AR is now widely used in China across over 800 schools [17], there has been no independent, rigorous evaluation of its impact on Chinese pupils to date. Most evaluations were conducted by the developers. Only one Chinese study [18] could be found. This was a small study (n = 29) conducted in a Chinese international school, using a single-group pre-post design with no counterfactuals. It reported that most English Language Learners (ELLs) showed gains in STAR reading scores and expressed positive attitudes toward reading. However, with no comparison group, it is not possible to attribute the positive effects to the intervention itself. The use of the STAR test, which is closely related to the content in the AR system, as an outcome measure also raises concerns about objectivity. Furthermore, the study took place in a relatively privileged context with native English teachers and a well-resourced library. Similar conditions are less likely to be found in state-funded Chinese schools.

There is thus a need for a more rigorous, contextually grounded evaluation of AR in Chinese public schools.

Prior evidence of AR

Although Accelerated Reader (AR) is widely used internationally and has been evaluated in numerous studies, its effectiveness in improving pupils' reading skills remains inconclusive. Most of these studies had methodological flaws. Paul, et al. [13] conducted a comprehensive study across 6,000 educational institutions in Tennessee, examining reading performance outcomes between schools implementing AR and those maintaining traditional instructional approaches. While their findings indicated enhanced performance among AR participants, several critical concerns emerged regarding study validity. The study was conducted by the developers of AR, raising concerns about conflict of interest. Although comparison schools were socioeconomically comparable to AR schools, schools that voluntarily adopted AR may have been more motivated or better resourced, introducing a selection bias that weakens the reliability of the reported effects. While Topping and Sanders [16] demonstrated that AR was positively correlated with pupils' reading performance, the study design

could not show causal effects. Ross, et al.'s [14] randomised control trial showed mixed results, with positive effects for younger children (K to grade 3), but no effects on the older children in grades 4 to 6. Again, the outcome measure was the AR-related STAR test.

More rigorous independent evaluations have yielded mixed findings. Gorard, Siddiqui and See's [12] evaluation of an efficacy study of AR among first-year secondary school pupils in England found that those who used AR made, on average, three months more progress than their peers who did not. Greater progress was noted among disadvantaged pupils (defined as those eligible for free school meals). Conversely, a more recent effectiveness trial by Sutherland, et al. [19], also conducted in England, found no effect on reading outcomes among Key Stage 2 (ages 9 to 11) pupils.

Other weaker studies have reported marginal or even adverse effects of AR. Mathis [20] found no significant impact of AR on pupils' reading comprehension. This was a single group pre-post study involving only 30 pupils from two classes, thus limiting its internal validity. Similarly, Pavonetti, et al. [21] documented no changes in pupils' reading habits, measured by the number of books read. However, their cross-sectional design with retrospective grouping could only establish associations rather than demonstrate causation between AR usage and reading frequency. A systematic review by the What Works Clearinghouse [22] synthesised findings from two investigations [23,24], including one with merely 32 participants. The review concluded that AR showed no impact on reading fluency, inconsistent effects on comprehension skills, and potentially positive influences on overall reading performance. Additionally, some researchers have raised concerns that AR's reward-based structure might undermine students' inherent motivation to read while restricting their book selection diversity [25,26].

In conclusion, although AR demonstrates potential in certain educational settings, the evidence base regarding its effectiveness remains inconclusive. Design limitations across existing research, including correlational approaches, insufficient sample sizes, and reliance on program-aligned assessments (STAR test), prevent definitive conclusions about AR's efficacy. Additionally, the predominance of research from English-speaking nations leaves AR's effectiveness in non-Western educational contexts, particularly China, largely unexamined.

- This research, therefore, aims to assess the impact of AR on Chinese school pupils' reading performance and reading behaviours. The research questions are:What impact does AR have on Chinese school pupils' reading performance?

- What impact does AR have on Chinese school pupils' reading behaviours (reading attitudes and reading habits)?

- Is there a correlation between reading behaviours and reading performance?

The intervention (Accelerated Reader)

Accelerated Reader (AR) is a computer-based reading programme designed to foster independent English reading habits among pupils. The programme is grounded in several learning and pedagogical theories from educational psychology and motivational research. A central principle of AR is matching pupils with books at their appropriate reading level, ensuring texts are sufficiently challenging to promote learning whilst avoiding excessive difficulty that might discourage engagement [27]. This aligns with Vygotsky's theory of Zone of Proximal Development (ZPD) [28]. Books within the AR system are therefore categorised using the Advantage/TASA Open Standard (ATOS) reading formula, whilst pupils complete the Standardised Test for the Assessment of Reading (STAR) to establish their appropriate reading level. It is important to note that AR does not provide books. In this pilot study, experimental classes were provided with approximately 215 English picture books by the researcher, covering a wide range of reading levels to ensure compatibility with the AR quiz database. This is something that schools wishing to use AR may need to consider.

Following completion of each book, pupils undertake a brief comprehension quiz. These quizzes generate TOPS reports (The Opportunity to Praise Pupils), delivering immediate performance feedback to each pupil. The AR system operates on motivational principles, whereby pupils accumulate reading points and receive rewards for meeting reading targets. Point allocation is determined by both the book's readability level and word count. AR is designed to encourage regular book reading that increases pupil's ZPD by stretching their reading abilities. An Interactive Reading Dashboard tracks each pupil's reading activity to help teachers monitor reading progress, and adjust pupils' reading goals to match their ZPD as they progress.

Prior to delivery, teachers received four hours of online training. The training covered STAR Reading, goal setting, quiz administration, and report interpretation. A WeChat group, which included trainers, IT staff, and customer service representatives, was set up to provide continuous support throughout the project.

Implementation of AR sessions occurred over 13 weeks in School A and 12 weeks in School B, with weekly sessions lasting approximately 40 minutes each. These sessions were integrated into regular English lessons. However, as pupils became more proficient at completing quizzes, session duration occasionally reduced to 20 minutes. Whilst AR recommends a minimum of 25 to 30 minutes daily independent reading, the actual number of books read and quizzes completed varies among individual pupils.

Methods

Trial design

The pilot was a two-armed cluster Randomised Control Trial (RCT), where classes were randomised within schools to either the AR intervention or a business-as-usual control group.

This approach helped minimise the risk of post-allocation demoralisation, as all schools participated as intervention schools. Randomisation ensured that the groups being compared were equivalent in terms of observed and unobserved characteristics and any differences in the performance could be attributed to the AR programme. The cluster RCT was chosen due to the logistical challenges of randomising individual pupils within class because of space and timetabling constraints.

Although the study was not structured as a wait-list trial, schools were informed that control classes could use the books provided for the AR group. This helped maintain school commitment to the programme.

The participants

The participants were Year 5 (age 10 to 11) and Year 6 (age 11 to 12) pupils from two public schools in the Anhui and Sichuan province in China. These schools were recruited through a combination of personal contacts, educational conferences and events, and outreach via Chinese social media platforms such as WeChat and Little Red Note, where a digital poster was used to introduce the project. Eight schools initially expressed interest, but most were not eligible as they had no access to computers or iPads, did not have capacity to establish a library of AR-related books, and no teachers available to attend AR training. In addition, AR is a paid programme requiring a subscription fee of approximately £26 per pupil, which posed a recruitment challenge for state-funded schools. For this pilot, the programme was offered free of charge by the researcher, with a discounted licence provided by Renaissance Learning, the developer. For this reason, the number of schools that could take part was small. Eventually, only two schools that met the eligibility criteria and were willing to take part in the randomised controlled trial were included in the study.

Years 5 and 6 pupils were selected because they were not in high-stakes exam years, making it more feasible for schools to accommodate the intervention. The younger Year 3 (ages 8–9) pupils were not considered as many were still developing basic decoding skills and not yet able to read English texts independently whereas Years 5 and 6 pupils had at least two years of English instruction and had developed stronger foundational reading skills, making them more suitable for the project. All the Year 5 pupils (n = 8 classes; 369 pupils) came from School A, while the Year 6 pupils (n = 3 classes; 159 pupils) were from School B. The total number of pupils was 528.

Randomisation

Randomisation was at the class level within schools. Because the headteacher in School A only allowed two classes to be included in the intervention group (due to lack of computer rooms), these were randomly selected from among all eligible classes. In contrast, all the three Year 6 classes in School B were included in the randomisation.

Due to the differences in year levels between the two schools, randomisation was conducted within each school at the class level, using pupils' English test scores from the previous

term as the basis for stratification. As there may be systematic differences between classes, randomisation by classes risked assigning high-attaining classes to one group. To mitigate this, classes were first ranked by their average English scores. In School A (with eight Year 5 classes), the top-ranked class was paired with the bottom-ranked class, the second-highest with the second-lowest, and so on, forming four balanced pairs. As the school could only accommodate two classes to receive the intervention, one pair (2 classes) was then randomly selected and assigned to the intervention group, while the remaining six classes served as the control group. In School B, which had three Year 6 classes, the highest- and lowest-performing classes were paired, with the remaining middle-performing class left unpaired. A random draw assigned the paired classes to the intervention group (2 classes), and the unpaired class to the control group (1 class). This stepwise, stratified approach was designed to maintain baseline balance. The randomisation results are summarised in Table 1.

In total, four classes (n = 195) were assigned to the intervention group, and seven (n = 333) to the control group, which continued with their usual lessons. Randomisation took place before pre-test during the summer holiday to allow time for teacher training and setting up of the intervention.

Although the intervention and control groups were unequal in size, this is methodologically acceptable in cluster trials, particularly when constrained by programme costs or limited institutional capacity, such as staffing or access to facilities (e.g., computers and computer rooms). Unequal allocation can preserve cost-efficiency and maintain adequate statistical power if total sample size remains sufficiently large [29].

However, such designs may result in baseline imbalances that warrant consideration during analysis. Table 2 shows that the gender and ethnic distributions of the pupils are broadly comparable, but there is a notable year group imbalance with 84% of control pupils in Year 5, compared to 46.2% in the intervention group, resulting in a higher average age in the intervention group. This difference may act as a confounding factor, as older pupils could have more English exposure or greater cognitive maturity. To account for baseline differences, analyses were performed using gain scores from pre-test to post-test.

Outcome measures

The primary outcome was pupils' English reading performance, which was assessed using the Cambridge Young Learner English (YLE) Test and the National Assessment Programme – Literacy and Numeracy (NAPLAN). Both the pre- and post-tests were administered online and consisted of 18 questions completed within 35 minutes, along with a reading behaviour survey.

The reading component included three texts that focused solely on comprehension skills. Text 1 was taken from the A1 Movers test, a Cambridge English qualification test [30] for young learners. It evaluated fundamental vocabulary skills through tasks requiring pupils to match words with corresponding pictures or phrases, establishing a baseline of English language competency. Texts 2 and 3 were drawn from Year 3 NAPLAN tests from 2012, 2015, and 2016 [31], as these materials closely aligned with the Chinese English curriculum for Years 5 and 6 [32]. Text 2 measured basic information retrieval skills, while Text 3 assessed higher-order comprehension through tasks involving the integration and interpretation of textual information. To ensure suitability, the difficulty level of the texts was analysed using the Flesch Reading Ease formula.

The secondary outcome was pupils' reading behaviour, which measured pupils' reading habits and reading attitudes. This was assessed using items from the 2018 Programme for International Student Assessment (PISA) [33], the 2021 Progress in International Reading Literacy Study (PIRLS) [6], and the Heathington Attitude Scale [34]. The questionnaire consisted of 12 items in total, including six items for reading habits and six for reading attitudes. For analyses, pupils' responses to the six items for each of the construct was totalled and a mean score obtained for reading habits and another for reading attitudes.

Analyses

Primary analysis: The primary analysis employed an intention-to-treat (ITT) approach, where all pupils were analysed according to their original random assignment. ITT maintains the original randomisation, thus minimising selection bias as all participants are analysed in the group to which they were assigned, regardless of whether they received the intervention or not. This helps mitigate potential biases from non-compliance or dropouts, maintaining the balance achieved by randomisation and providing an unbiased estimate of the treatment's impact in real-world conditions [35]. It also reduces the risk of distortion if those dropped out were in some way different to those who stayed.

The impact of AR was estimated using Hedge's g effect size (*ES*), calculated as the difference in the gain scores between the intervention and control groups, divided by the overall standard deviation (*SD*). Gain scores were used as there was initial imbalance at pre-test.

Table 1: Class-level randomisation to AR and control by school and year level.

School name	Year level	Experimental group	Control group
School A	5	2 classes (90 pupils)	6 classes (279 pupils)
School B	6	2 classes (105 pupils)	1 class (54 pupils)
Total	5 & 6	4 classes (195 pupils)	7 classes (333 pupils)

Table 2: Demographic characteristics of pupils in the AR and control groups.

Characteristics of pupils	AR (%) (n = 195)	Control (%) (n = 333)	Differences (%)
Boys	49.2	53.2	-4.0
Minority	4.6	3.3	-1.3
Year 5	46.2	83.8	-37.6
Year 6	53.8	16.2	37.6
Age	11.74	11.27	0.47

Note that we do not use significance tests and confidence intervals as they rely on assumptions of full randomisation and no missing values [36,37], which do not hold in our study. And even if these conditions were met, they would still not be appropriate because they only tell us the probability of observing the results we get assuming that there is no difference between the groups [38-40]. What we are interested to know is whether there is, indeed, any difference between groups. Under such conditions, relying on significance testing would be inappropriate and potentially misleading.

Instead, to assess the security of the findings, that is whether it could occur by chance or influenced by bias due to attrition, we calculated the 'number needed to disturb' the finding [41]. This indicates the number of counterfactual cases that would be needed to reverse or alter the observed result. The larger the NNTD, the more stable the finding.

NNTD is calculated as the "effect" size (*ES*) multiplied by the number of cases in the smallest group in the comparison (i.e., the number of cases included in either the control or treatment group, whichever is smaller).

Since missing responses in reading habits and attitudes did not constitute attrition, and to avoid losing cases due to partial missing value, missing values were replaced with the mean of the total scale (sum of six items divided by six). This method avoids item-level bias and maintains scale-level variance [42], ensuring most participants remained in the analysis.

Additional analysis: While the impact evaluation and sensitivity analysis established the overall effects and their reliability, multiple linear regression analysis was used to determine how much the observed differences in reading performance and reading behaviours were due to the intervention or other confounding factors [43].

In these three regression analyses, the post-test or post-survey scores were used as the dependent variable, with pre-test scores entered as covariates. Independent variables were added in three blocks, sequentially in chronological order. The first block included pupil demographic factors such as gender, ethnicity, year level and age. These were included first as these are factors that are not malleable. The second block included relevant pre-test or pre-survey scores. The final block added the treatment group status, to assess the additional contribution of intervention participation.

As several predictors were included, the adjusted R-squared was calculated to produce a predictive model. The increase in variance explained at each step shows how much more these variables add to predicting pupils' reading outcomes.

Since not all pupils received the recommended number of sessions, ITT analysis may underestimate the true effects. To explore this further, we first examined fidelity to treatment by analysing the number of sessions each pupil completed (as recorded on the AR dashboard) and its correlation with their English test scores.

To estimate the causal effect of AR among those who complied with the recommended usage, we conducted a Complier Average Causal Effect (CACE) analysis [44]. Essentially, it is a comparison of what actually happens with what might have happened [45]. The complier average causal effect is estimated using known information about the treatment performance and the assumptions that because of randomisation the proportion of compliers in the control would be the same as for treatment, and the average performance for those in the control group who did not comply would be the same as the performance of non-compliers in the treatment group (Cell D, Table 3).

Given that we know the overall results for both groups (Cells F & K) and the data for those in the treatment group who complied and who did not comply (Cells A to D), we can calculate the average performance for those in the control group who would have complied if given the treatment (x). The proportion in treatment group who complied is assumed to be A/E:

- Number in control group who complied (Cell G) will be A/E*J

- Number of non-compliers in control group (Cell H) will be J-G

- The average performance for compliers in the control group (x) is calculated thus:

x = ((J*K) – (H*I))/G

In this study, a complier was defined as a pupil who completed AR quiz at least once a week for 12 or 13 weeks.

Building on the previous analyses of impact and fidelity, the path analysis is used to explore how AR usage related to changes in pupils' reading behaviours (habits and attitudes) and subsequent reading performance, providing insights into the "black box" of the intervention's effectiveness (Figure 1).

The path analysis was constructed using two regression models: one assessed the direct effect of the intervention on reading behaviours and the reading performance, and the second examined the effect of the mediating factors (reading behaviours) on the reading performance. This approach provided insight into the underlying mechanisms driving the relationship between AR and English reading performance. The standardised regression coefficient (beta) or path coefficient indicated the strength and direction of the relationship between the variables within the model.

The outcome variables were gain scores. This choice avoided potential biases from pre-test scores and baseline differences among pupils, allowing clearer examination of

Table 3: Estimation of complier average causal effect.

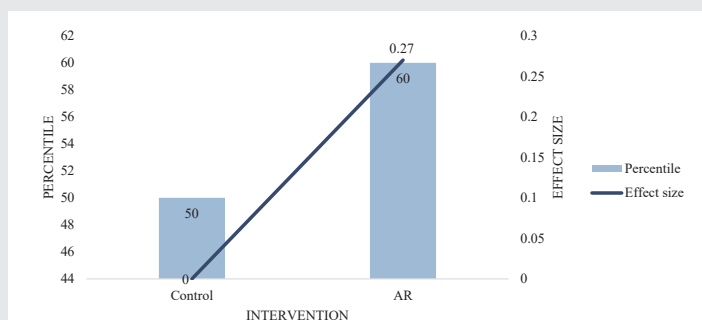|	Compliers		Non-compliers		Overall	
	N who completed AR quizzes	Mean	*N* who did not complete AR quizzes	Mean	*N*	Mean
Treatment	A	B	C	D	E	F
Control	G	x	H	I	J	K

051

Figure 2: Comparing English reading performance by effect Size and percentile growth.

the relationships between AR usage, reading behaviours, and performance. Background variables (gender, ethnicity, year level, and age) were excluded from the path analysis as these were already accounted for in the earlier multiple linear regression. Including these factors could have obscured the direct and mediated effects of AR. Adjusted R-squared values were calculated to evaluate the proportion of variance explained by the predictors in each model. This measure ensured that the models provided an accurate and parsimonious explanation of the observed performance.

Process evaluation

The purpose of the process evaluation was to assess both the fidelity (that is if the programme was delivered as intended) and the quality of implementation of the intervention [46,47]. It provides a better understanding of the underlying mechanism that may explain how and why the intervention worked, if it was successful. And if the programme was found to be ineffective, it could explain whether the programme is intrinsically ineffective, or the teachers were not implementing it in the way they should. It also helped to identify potential limitations [48] and capture unintended consequences if any.

The process evaluation included classroom observations and informal interviews with pupils and teachers. Another important source of data came from AR's interactive dashboard, which tracked pupils' reading, their reading level, the number of books read, and the number of quizzes completed. This provides information on the fidelity to intervention, e.g., whether pupils have completed the required number of quizzes.

School visits were arranged via headteachers to observe programme delivery, pupils' reactions to the programme and teachers' ability to use the resources, while noting any emerging changes in classroom practice or pupil learning. There was no structured interview schedule as such as the intention was to capture pupils' and teachers' perceptions of the intervention. We had intentionally kept the informal conversations open so that we were not influenced by what we thought was important, but what the participants thought was important, and to allow the evidence to speak for itself without prejudice. So, all information was potentially relevant here.

During the school visits, we conducted non-intrusive classroom observations. These included learning walks, during which we examined pupils' work and listened to their

spontaneous comments. Opportunities often arose to engage with pupils while they completed AR quizzes. In School A, a pupil focus group was organised to facilitate a casual discussion with pupils. Observations, pupils' feedback and informal remarks were documented through note-taking – no audio or video recordings were made.

In total four visits were made to two treatment schools to observe the process of implementation in 46 sessions. Individual interviews with teachers were delivered online at the end of the project, focusing on their perceptions of AR programme. Specifically, what they thought had contributed or would contribute to the success of the programme, and the barriers to effective delivery of the intervention.

Ethical considerations

Ethical approval for this study was granted by Durham University's Ethics Committee (Reference: EDU-2022-11-29T11_12_14-pcgm56) on 15 February 2023. Participation was entirely voluntary. An opt-out consent procedure was used: only parents who did not wish their children to take part were required to return a signed form. No personally identifiable information was collected, and all data were anonymised for the purposes of analysis and reporting. All procedures followed the ethical principles outlined in the Declaration of Helsinki (1975, revised 2013).

Findings

Impact of AR on pupils' reading performance: Since pupils in the AR group were already ahead at pre-test, it would not be fair to use post-test scores only. Therefore, gain scores were used to compare progress made between the groups. Table 4 shows that both groups performed worse at post- test than at pre-test, likely attributable to the greater difficulty of the post-test compared to the pre-test, but the AR group showed less decline than the control group, suggesting that the AR group performed better in comparison (ES = +0.27).

To express this effect in percentile growth (Figure 2), it means that the median pupils who ranked at the 50th percentile in the control group, if they had received AR, would now be ranked at 60th percentile, demonstrating the positive impact of the AR intervention [49].

While effect sizes reflect the magnitude of the difference, they do not account for study scale or robustness of the effects. To test the robustness of the finding, the Number Needed to Disturb (NNTD) was calculated. For this trial, the effect size was +0.27, and the smaller group (AR group) had 195 pupils, resulting in an NNTD of 53 (195 × 0.27). This means that 53 extreme negative scores would be needed to nullify or reverse the positive effect. Since there were no missing cases, the findings can be considered robust and secure.

Table 4: Comparing gain scores of English reading performance between AR and control.

	N	Pre-test	SD	Post-test	SD	Gains	SD	ES
AR	195	8.51	3.62	8.09	4.43	-0.42	4.56	
Control	333	8.14	3.51	6.58	3.38	-1.56	3.99	0.27
Total	528	8.27	3.55	7.14	3.86	-1.14	4.24	

Although the effect size suggested a positive impact of the AR programme on reading performance, the results of the regression analysis showed that pupils' year group alone explained 23% of the variance in post-test scores (Table 5). Gender and pre-test scores added a small amount to the prediction. The intervention adds nothing to the predictive accuracy of the model. This suggests that most of the differences in reading performance could be explain by pupils' year group.

Impact of AR on pupils' reading behaviours

As with the reading performance, the results for reading habit show that both groups made negative progress from pre-test to post-test, but in comparison, the decline was much bigger for the control group (Table 6). In other words, AR group performed better than control (ES = +0.15), particularly in areas like reading for pleasure, borrowing books, and extracurricular reading (see Appendix A for a full list of items for reading habits). The intervention may have mitigated against the negative effect seen in the control group. AR may have helped maintained motivation and routines, thus preventing the sharp decline in reading habits for children in this age group.

However, the regression analysis (Table 7) showed that the strongest predictor of reading habits is pupils pre-reading habits. Pupil's year group explained 3.4% of the variance initially. Adding pre-habits reading habits improved the accuracy of prediction by 9.3%. The intervention did not contribute to explaining the post-reading habits, suggesting that the intervention had no direct effect on pupils' reading habits.

For reading attitude, AR also showed a small positive effect (Table 8). Positive effects were observed in five out of six individual items. This is not surprising as AR requires pupils to read in their free time, and select books matched to their interests and reading levels. While AR may enhance pupils' enjoyment of reading, it did not improve their confidence in reading English in class (see Appendix B for a full list of items for reading attitudes). It is probably because the focus of AR focuses is on individual, self-paced reading, not public reading.

As with reading achievement and reading habits, AR did not explain the outcome once other background factors, such as prior attitudes and year group were controlled (Table 9). This suggests that the observed attitude gains may be shaped more by pupils' baseline characteristics than by AR itself.

Fidelity to implementation

Since not all pupils completed the required number of sessions or quizzes (between 12 and 13), a compliance analysis was performed to see if compliance made a difference to the performance. The average number of AR sessions completed by pupils in the treatment group was 16.81. However, there was substantial variation across schools. Of the treatment group, 102 learners did not achieve the minimum number of sessions recommended (13 sessions for School A and 12 sessions for School B), while 93 learners received the recommended lessons with a large percentage well in excess of this number. Several

Table 5: Regression model predicting the post-test performance.

Model	R	R^2	Adjusted R^2	R^2 Change
1 Year group	0.48	0.23	0.23	0.23
2 Year group & gender	0.49	0.24	0.23	0.01
3 Year group, gender & pre-test	0.53	0.28	0.27	0.04
4 Year group, gender, pre-test and treatment	0.53	0.28	0.27	0

Table 6: Comparing gain scores of English reading habits between AR and control groups.

	N	Pre-survey	SD	Post- survey	SD	Gain	SD	ES
AR	195	5.20	2.50	5.10	2.53	-0.10	2.96	
Control	333	5.33	2.22	4.84	2.61	-0.49	2.73	
Total	528	5.28	2.32	4.94	2.58	-0.34	2.82	0.14

Note: Reading habits represent the average gain across six items. For item-level results, see Appendix A.

Table 7: Regression model predicting post-reading habits.

Model	R	R^2	Adjusted R^2	R^2 Change
1 Year group	0.183	0.034	0.032	0.034
2 Year group and pre-reading habits	0.357	0.127	0.124	0.093
3 Year group, pre-reading habits and treatment	0.357	0.127	0.122	0

Table 8: Comparing gain scores of English reading attitudes between AR and control groups.

	N	Pre-survey	SD	Post- survey	SD	Gain	SD	ES
AR	195	6.62	1.95	5.94	1.91	-0.69	2.34	
Control	333	6.57	1.87	5.55	1.82	-1.01	2.18	0.15
Total	528	6.59	1.90	5.70	1.86	-0.89	2.24	

Note: Reading attitudes represent the average gain across six items. For item-level results, see Appendix B.

Table 9: Regression model predicting post-reading attitudes.

Model	R	R^2	Adjusted R^2	R^2 Change
1 Year group	0.223	0.050	0.048	0.050
2 Year group and pre-reading attitudes	0.343	0.118	0.114	0.068
3 Year group, pre-reading attitudes and treatment	0.344	0.118	0.113	0

reasons contributed to low dosage among some pupils. Frequent absences, limited access to labelled books, technical difficulties with the AR system, and low English proficiency were commonly reported. Some pupils also struggled with logging in or completing quizzes, while others rushed through the sessions due to classroom distractions or time constraints. Pupils with weaker English reading skills found it particularly hard to sustain reading and quiz participation without additional support.

Correlation analysis showed a small positive relationship (+0.22) between gain scores in English reading test and the number of sessions received. In other words, the more sessions a pupil received, the higher their English scores (and vice versa). It is difficult to say exactly what this means because pupils who missed sessions or who were regularly absent may face other issues which contribute to their performance (but which are not investigated within the scope of this project).

There may also be the increased likelihood of issues in relation to motivation or confidence for these pupils.

To account for variations in compliance with the recommended number of AR sessions, a CACE analysis was conducted, establishing a threshold of 13 AR quizzes for School A and 12 quizzes for School B. The analysis aimed to estimate the treatment effect in cases where some pupils in the treatment group did not meet the minimum dosage, as well as to the project outcome for control group pupils had they complied with the intervention.

From the treatment group (Table 10), 93 out of 195 pupils (48%) met the compliance threshold, completing the required number of quizzes. By assuming the same proportion of compliance in the control group, it was estimated that 160 out of 333 control pupils would have complied if they had been assigned to the intervention. The mean performance for non-compliers in both the control and treatment groups was calculated to be -1.76, allowing for an estimation of what compliant control pupils' scores might have been.

Using the overall standard deviation from Table 4 (SD = 4.24), the effect size based on compliers in the treatment group was calculated using the formula $((1.05-(-1.34))/4.24)$. The CACE analysis revealed a higher complier effect size of +0.56, compared to the overall headline effect size of +0.27 for the gain scores. These results suggest that when the AR programme is implemented as intended and pupils complete the recommended number of quizzes, the positive effects of the intervention will be stronger, indicating that AR intervention can be highly effective when implemented with fidelity, benefiting pupils who fully engage with the programme.

Does improving reading behaviours lead to improvement in reading performance?

To answer this question, a path analysis was conducted to examine the effects of the mediating factors (i.e., reading habits and reading attitudes) on reading performance.

The results of the path analysis (Figure 3) showed that AR had a direct and positive effect on reading performance as indicated by the path coefficient ($ß$ = 0.13). AR also had a positive direct effect on reading habits and reading attitudes ($ß$ = 0.07 and $ß$ = 0.07, respectively). However, reading habits had a small negative effect ($ß$ = -0.01), while reading attitudes had a small positive effect ($ß$ = 0.10) on reading achievement.

- The combined effect of the intervention on reading outcome through the mediating factors was its direct effect plus indirect effecReading habits: $0.07 × (-0.01) = -0.0007$

Table 10: CACE analysis based on 12-13+ sessions and English reading performance gain scores.

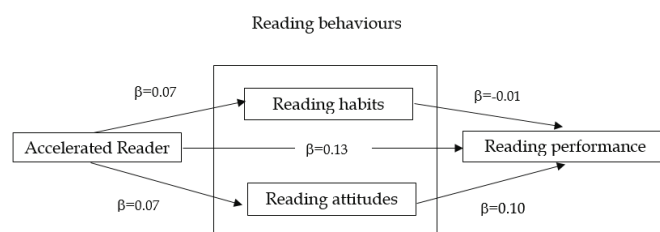|	Completed		Not Completed		Overall	
	N	Mean	N	Mean	N	Mean
AR	93	1.05	102	-1.76	195	-0.42
Control	160	-1.34	173	-1.76	333	-1.56

Figure 3: Results of the path analysis.

- Reading attitudes: $0.07 × 0.10 = 0.007$

- Direct effect = 0.13

- Indirect effect = 0.0063

- Total combined effect ≈ 0.14

This indicates that AR might be multifaceted, benefitting some aspects of learning behaviours (reading attitudes), but not reading habits, suggesting that perhaps reading habits is not an essential mediator (or even counterproductive) for the intervention to be effective.

Limitation

While the trial was well designed, several limitations beyond the evaluators' control might have adversely influenced the strength of the evidence. The biggest limitation was the variation in implementation fidelity between the two schools. Although teachers received initial training, some required more ongoing support to deliver AR programme effectively. Inconsistent use of key AR features, such as the dashboard for monitoring pupils' progress and the frequency of AR sessions, likely affected the quality of delivery. CACE analysis revealed that consistent and fidelity to implementation is essential for AR to have an effect on pupils' reading and behavioural outcomes.

There was also evidence of diffusion as the control pupils also had access to other online educational platforms, such as the 17zuoye. These platforms provided additional opportunities for English reading practice outside of AR. Although efforts were made to control for external factors, the exposure to supplementary learning resources in the control group could have muted the effects. The key strength of an RCT is the clean comparison between groups. Diffusion blurs this comparison, threatening the study's internal validity [50].

Due to time-tabling constraints and the lack of staff capacity, randomisation had to be at the class level to minimise disruption to regular lessons. The use of cluster randomisation, where entire classes were assigned to either the treatment or control group had significantly reduced the sample size and thus the statistical power of the study. This issue was compounded by the request of one school to allow only two of their eight Year 5 classes to receive the intervention. This led to an imbalance in the number of control and intervention classes, with more classes being in the control group. While necessary for practical reasons, this design had potentially

impacted the comparability between the treatment and control groups [51].

The short duration of the intervention (12-13 intervention weeks) also means that the effects of the intervention may not have time to manifest itself. As reading attitude and reading habits take time to develop, a longer duration may be necessary. Developing sustained reading habits may require a longer period to influence academic performance, particularly in second language contexts where cognitive processing demands are higher [52].

Discussion and implications

This pilot trial was the first RCT evaluation of AR within the Chinese contexts, and the results showed that AR has some promise in supporting Chinese primary pupils' English reading performance and reading behaviours. However, regression analyses suggest that confounding factors (especially year level and baseline test or survey) may influenced the observed effects. These findings require further investigation using larger, more balanced samples. Nevertheless, several important lessons emerged from this pilot study that could inform future AR evaluations.

Close monitoring of the programme

The process evaluation showed considerable variations among pupils regarding the number of sessions completed. CACE analysis showed that the number sessions completed mattered. Pupils who completed more sessions or quizzes completed, made slightly more progress. This suggests that the intervention can lead to a better performance if implemented with fidelity. To ensure that children complied with the minimum required number of sessions, closer monitoring and supervision of teachers are necessary.

However, it is also possible that the kind of pupils who completed more quizzes would make more progress anyway. Further analysis to disentangle the effect of pupil characteristics (e.g., motivation, their prior attitude and habits) and the number of sessions conducted would provide better insight into this.

Future implementations may benefit from on-site support during setup to ensure that teachers and school leaders are proficient in using the AR dashboard to set reading goals, track pupils' progress, and effectively implement the AR reward system, and are clear about their roles and responsibilities. This helps ensure that the programme is implemented as intended.

Practical considerations in Chinese contexts

As AR is a paid programme requiring subscription licence (£26 per pupil), which require access to computers or iPads and a suite of library books, public schools in China considering adopting AR need to consider these costs. The large class sizes in most Chinese schools may make the adoption of AR challenging. This pilot study shows that it is possible, but it needed a lot of support from the school leaders.

Long-term effects and other subjects

As the pilot trial was conducted in less than one term, it would be valuable to investigate whether a longer duration, such as 20 weeks or more, might produce stronger effects on reading performance and reading behaviours. **Studies with durations of 20 weeks or longer are common in AR evaluations in English-speaking countries** [12,14,19,53]. This is especially relevant as improving attitudes and habits often take longer to shift.

In most research studies, the intervention ends when the researchers leave the field. It would be interesting to see if the schools continue with using AR after the trial. This would demonstrate schools' commitment and belief in the efficacy of the intervention. A longitudinal study, following children a year after the trial, could also be conducted to look at the long-term impact of AR. Prior research indicates that strong literacy skills are critical for subsequent academic performance [54] and achievement in other subjects [55]. Thus, it would be worthwhile to explore whether AR has enduring positive effects on pupil performance over time and any spillover effects on other subjects.

(**Appendix**)

References

1. Qi GY. The importance of English in primary school education in China: perceptions of students. Multilingual Education. 2016;6(1):1-18. Available from: http://dx.doi.org/10.1186/s13616-016-0026-0

2. Ministry of Education. Number of students in primary school by types. Ministry of Education of China; 2023. Available from: http://en.moe.gov.cn/documents/statistics/2022/national/202401/t20240110_1099490.html

3. Ali Z, Palpanadan ST, Asad MM, Churi P, Namaziandost E. Reading approaches practiced in EFL classrooms: a narrative review and research agenda. Asian-Pacific Journal of Second and Foreign Language Education. 2022;7(1):28. Available from: https://doi.org/10.1186/s40862-022-00155-4

4. Liang Y, Nian OS. The Challenges Brought by China's New English Textbooks for Senior High Students. Platform: A Journal of Management and Humanities. 2023;6(2):2-15.

5. Mullis I, Von Davier M, Foy P, Fishbein B, Reynolds K, Wry E. PIRLS 2021 International results in reading. 2023. Available from: https://doi.org/10.6017/lse.tpisc.tr2103.kb5342

6. Mullis IVS, Martin MO. PIRLS 2021 context questionnaire frameworks. TIMSS & PIRLS International Study Center; 2019. Available from: https://pirls2021.org/frameworks/wp-content/uploads/sites/2/2019/04/P21_FW_Ch2_Questionnaires.pdf

7. Locher F, Pfost M. The relation between time spent reading and reading comprehension throughout the life course. Journal of Research in Reading. 2020;43(1):57-77. Available from: https://doi.org/10.1111/1467-9817.12289

8. Van Bergen E, Vasalampi K, Torppa M. How are practice and performance related? Development of reading from age 5 to 15. Reading Research Quarterly. 2020;56(3):415-34. Available from: https://doi.org/10.1002/rrq.309

9. McKenna MC, Kear DJ, Ellsworth RA. Children's Attitudes toward Reading: A National Survey. Reading Research Quarterly. 1995;30(4):934-56. Available from: https://doi.org/10.2307/748205

10. Petscher Y. A meta-analysis of the relationship between student attitudes towards reading and achievement in reading. Journal of Research in

Reading. 2009;33(4):335–55. Available from: https://doi.org/10.1111/j.1467-9817.2009.01418.x

11. Tunnell MO, Calder JE, Justen JE, Phaup ES. Attitudes of young readers. Reading Improvement. 1991;28(4):237.

12. Gorard S, Siddiqui N, See BH. Accelerated Reader: Evaluation report and executive summary. Education Endowment Foundation; 2015. Available from: https://files.eric.ed.gov/fulltext/ED581101.pdf

13. Paul TD, VanderZee D, Rue R, Swanson S. Impact of Accelerated Reader on overall academic achievement and school attendance. Paper presented at: National Reading Research Center Conference, Literacy and Technology for the 21st Century; 1996 Oct; Atlanta, GA.

14. Ross S, Nunnery J, Goldfeder E. A randomized experiment on the effects of Accelerated Reader/Reading Renaissance in an urban school district: Preliminary evaluation report (Research report). The University of Memphis; 2004.

15. Ross SM, Nunnery JA. The effect of school Renaissance on student achievement in two Mississippi school districts. University of Memphis, Center for Research in Education Policy; 2005. Available from: https://files.eric.ed.gov/fulltext/ED484275.pdf

16. Topping K, Sanders W. Teacher effectiveness and computer assessment of reading relating value added and learning information system data. School Effectiveness and School Improvement. 2000;11(3):305–37. Available from: https://doi.org/10.1076/0924-3453(200009)11:3;1-g;ft305

17. Renaissance Learning. About us. Renaissance company; 2024. Available from: https://renaissance.cn

18. Zhang H. The Impacts of Accelerated Reader on English Language Learners [dissertation]. [Ann Arbor]: ProQuest; 2023. Order No. 31746816. Available from: https://www.proquest.com/docview/3132885167

19. Sutherland A, Broeks M, Ilie S, Sim M, Krapels J, Brown ER, Belanger J. Accelerated Reader evaluation report (Research report). Education Endowment Foundation; 2021. Available from: https://d2tic4wvo1iusb.cloudfront.net/production/documents/pages/projects/Accelerated_Reader_-_final.pdf?v=1748512496

20. Mathis D. The effect of the Accelerated Reader program on reading comprehension (Research report). Educational Resources Information Center (ERIC); 1996. Available from: https://files.eric.ed.gov/fulltext/ED398555.pdf

21. Pavonetti LM, Brimmer KM, Cipielewski JF. Accelerated Reader: What are the lasting effects on the habits of middle school students exposed to Accelerated Reader in elementary grades? Journal of Adolescent & Adult Literacy. 2002;46(4):300–11.

22. What Works Clearinghouse. Accelerated Reader: WWC intervention. U.S. Department of Education; 2008. Available from: https://ies.ed.gov/ncee/wwc/Docs/InterventionReports/wwc_accelreader_101408.pdf

23. Bullock JC. Effects of the Accelerated Reader on reading performance of third, fourth, and fifth-grade students in one western Oregon elementary school [doctoral dissertation]. University of Oregon; 2005. (Publication No. 3181085). Available from: https://www.proquest.com/dissertations-theses/effects-accelerated-reader-on-reading-performance/docview/305405743/se-2

24. Shannon LC, Styers MK, Siceloff ER. A final report for the evaluation of Renaissance Learning's Accelerated Reader Program (Research report). Magnolia Consulting; 2010.

25. Biggers D. The argument against Accelerated Reader. Journal of Adolescent & Adult Literacy. 2001;45:72–5. Available from: https://dianedalenberg.wordpress.com/wp-content/uploads/2012/05/argument-against-ar.pdf

26. Stevenson JM, Camarata JW. Imposters in whole language clothing: Undressing the Accelerated Reader program. Talking Points. 2000;11:8–11. Available from: https://www.learntechlib.org/p/93593/

27. Renaissance Learning. Zone of proximal development. Accelerated Reader; 2024. Available from: https://arhelp.renaissance.com/hc/en-us/articles/12746104128027-Zone-of-Proximal-Development

28. Vygotsky LS. Interaction between Learning and Development. In: Cole M, John-Steiner V, Scribner S, Souberman E, editors. Mind and Society: The Development of Higher Psychological Processes. Cambridge, MA: Harvard University Press; 1978;79-91. Available from: https://www.scirp.org/reference/referencespapers?referenceid=1929734

29. Torgerson DJ, Torgerson CJ. Designing randomised trials in health, education and the social sciences: An introduction. Palgrave Macmillan; 2008. Available from: http://dx.doi.org/10.1057/9780230583993

30. Cambridge English Qualification. A1 Movers preparation: Sample tests; 2023. Available from: https://www.cambridgeenglish.org/exams-and-tests/movers/preparation/

31. Australian Curriculum, Assessment and Reporting Authority. NAPLAN 2012–2016 test papers and answers. 2023. Available from: https://acara.edu.au/assessment/naplan/naplan-2012-2016-test-papers

32. Ministry of Education of the People's Republic of China. English curriculum standards for compulsory education. Beijing: Beijing Normal University; 2022. Available from: annex_4_-_cnec_translated_version_final.pdf

33. OECD. PISA 2018: Insights and interpretations. Paris: OECD; 2019. Available from: https://shorturl.at/N3YR7

34. Heathington BS. The development of scales to measure attitudes towards reading [dissertation]. Knoxville (TN): University of Tennessee; 1975. Available from: https://trace.tennessee.edu/utk_graddiss/3089/

35. Wertz RT. Intention to treat: Once randomized, always analyzed. Clin Aphasiol. 1995;23:57-64. Available from: http://eprints-prod-05.library.pitt.edu/188/1/23-05.pdf

36. Berk RA, Freedman DA. Statistical assumptions as empirical commitments. In: Blomberg TG, Cohen S, editors. Law, punishment, and social control: Essays in honor of Sheldon Messinger. 2nd ed. Aldine de Gruyter; 2003. p. 235–54. Available from: https://www.stat.berkeley.edu/~census/berk2.pdf

37. Lipsey MW, Puzio K, Yun C, Hebert MA, Steinka-Fry K, Cole MW, et al. Translating the statistical representation of the effects of education interventions into more readily interpretable forms (NCSER 2013-3000). 2012. Available from: https://shorturl.at/8REiF

38. Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. R Soc Open Sci. 2014;1(3):140216. Available from: https://doi.org/10.1098/rsos.140216

39. Colquhoun D. The problem with p-values. 2016. Available from: https://aeon.co/essays/it-s-time-for-science-to-abandon-the-term-statistically-significant

40. Gorard S. Damaging Real Lives through Obstinacy: Re-Emphasising Why Significance Testing is Wrong. Sociol Res Online. 2016;21(1):102–15. Available from: https://doi.org/10.5153/sro.3857

41. Gorard S. Education policy. Policy Press eBooks; 2018. Available from: https://doi.org/10.1332/policypress/9781447342144.001.0001

42. Rombach I, Gray AM, Jenkinson C, Murray DW, Rivero-Arias O. Multiple imputation for patient reported outcome measures in randomised controlled trials: Advantages and disadvantages of imputing at the item, subscale or composite score level. BMC Med Res Methodol. 2018;18:87. Available from: https://doi.org/10.1186/s12874-018-0542-6

43. Gorard S. How to make sense of statistics: Everything you need to know about using numbers in social science. SAGE Publications; 2021. Available from: https://methods.sagepub.com/book/mono/preview/how-to-make-sense-of-statistics.pdf

44. Dunn G, Maracy M, Tomenson B. Estimating treatment effects from randomized clinical trials with noncompliance and loss to follow-up: The role of instrumental variable methods. Stat Methods Med Res. 2005;14(4):369–95. Available from: https://doi.org/10.1191/0962280205sm403oa

45. See BH, Gorard S, Lu B, Dong L, Siddiqui N. Is technology always helpful? A critical review of the impact on learning outcomes of education technology in supporting formative assessment in schools. Res Pap Educ. 2022;37(6):1064–96. Available from: https://doi.org/10.1080/02671522.2021.1907778

46. Carroll C, Patterson M, Wood S, Booth A, Rick J, Balain S. A conceptual framework for implementation fidelity. Implement Sci. 2007;2(1). Available from: https://doi.org/10.1186/1748-5908-2-40

47. Montgomery P, Underhill K, Gardner F, Operario D, Mayo-Wilson E. The Oxford Implementation Index: a new tool for incorporating implementation data into systematic reviews and meta-analyses. J Clin Epidemiol. 2013;66(8):874–82. Available from: https://doi.org/10.1016/j.jclinepi.2013.03.006

48. Steckler AB, Linnan L, Israel BA. Process evaluation for public health interventions and research. 2002. Available from: https://psycnet.apa.org/record/2003-02384-000

49. Baird MD, Pane JF. Translating standardized effects of education programs into more interpretable metrics. Educ Res. 2019;48(4):217–28. Available from: https://doi.org/10.3102/0013189X19848729

50. Shadish WR, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin; 2001.

51. Donner A, Klar N. Design and analysis of cluster randomization trials in health research. Oxford University Press; 2000.

52. Linck JA, Osthus P, Koeth JT, Bunting MF. Working memory and second language comprehension and production: A meta-analysis. Psychon Bull Rev. 2013;21(4):861–83. Available from: https://doi.org/10.3758/s13423-013-0565-2

53. Shannon LC, Styers MK, Wilkerson SB, Peery E. Computer-Assisted Learning in Elementary Reading: a randomized control trial. Comput Sch. 2015;32(1):20–34. Available from: https://doi.org/10.1080/07380569.2014.969159

54. Cunningham AE, Stanovich KE. Early reading acquisition and its relation to reading experience and ability 10 years later. Dev Psychol. 1997;33(6):934. Available from: https://psycnet.apa.org/doi/10.1037/0012-1649.33.6.934

55. Hall SS, Kowalski R, Paterson KB, Basran J, Filik R, Maltby J. Local text cohesion, reading ability and individual science aspirations: Key factors influencing comprehension in science classes. Br Educ Res J. 2014;41(1):122–42. Available from: https://doi.org/10.1002/berj.3134ts: