

Effects of a Blended Deep Learning Module on Deep Reading Ability and English Reading Achievement among Chinese Primary School Pupils: A Quasi-Experimental Study

Yuxuan Wen, Nur Jahan Ahmad*

School of Educational Studies, Universiti Sains Malaysia, 11800 Gelugor, Pulau Pinang, Malaysia.

*Corresponding Author

DOI: <https://doi.org/10.47772/IJRISS.2026.100801012>

Received: 03 September 2026; Accepted: 16 September 2026; Published: 23 September 2026

ABSTRACT

Primary English reading instruction in China remains largely oriented toward examination performance and literal comprehension, leaving pupils with few opportunities to infer, evaluate, and respond creatively to texts. This study examined the effects of a Blended Deep Learning (BDL) module on sixth-grade pupils' deep reading ability and English reading achievement. The module treats deep learning as the pedagogical aim and the blended structure as the delivery mechanism, and it was developed by integrating the ASSURE instructional design model with the Deeper Learning Cycle across three phases: pre-class online preparation, in-class face-to-face learning, and post-class online consolidation. A quasi-experimental nonequivalent control group pretest-posttest design was employed with two intact Grade 6 classes ($N = 60$) in a public primary school in Jiangsu Province, China. The treatment group learned through the BDL module for eight weeks while the control group received conventional instruction, and both groups were taught by the same teacher. Deep reading ability was measured with the Deep Reading Ability Questionnaire, covering inferential, critical, and creative reading skills, and achievement with an English Reading Test. MANOVA was used to examine post-test deep reading outcomes after baseline equivalence had been established, and ANCOVA was used to examine English reading achievement while controlling for pre-test achievement. The analyses showed statistically significant differences between the groups across the combined deep reading dimensions and, after adjustment for pre-test scores, in English reading achievement. The findings suggest that a purposefully integrated blended design can support higher-order reading skills and reading achievement in primary EFL classrooms, while the single-school quasi-experimental design requires cautious interpretation and further replication across larger and more diverse samples.

Keywords: Blended deep learning, Deep reading ability, English reading achievement, Primary school pupils, Quasi-experimental design

INTRODUCTION

Reading is the language skill through which most of a primary pupil's English learning is consolidated, yet the way reading is taught in many Chinese primary classrooms leaves little room for the kind of thinking that turns decoding into understanding. Under pressure from high-stakes assessment, lessons commonly concentrate on locating stated information and rehearsing question types, so that comprehension is treated as an end product to be checked rather than a process to be developed. International reading scholarship has for some time argued that comprehension worthy of the name requires readers to go beyond the literal: to infer what a text implies, to weigh the author's claims and stance, and to generate their own responses and alternatives (Wolf & Barzillai, 2009; LaRusso et al., 2016). Chinese curriculum reform has moved in the same direction. The Compulsory Education English Curriculum Standards list "thinking quality" among the core competencies that primary English is expected to develop, alongside language ability, cultural awareness, and learning ability. What has been slower to emerge is classroom pedagogy that operationalises this aspiration for young learners in a systematic and assessable way.

A parallel development has been the rapid institutionalisation of blended learning in Chinese basic education, accelerated by provincial informatisation plans and by the widespread deployment of educational cloud platforms, interactive whiteboards, and pupil tablets (Geng et al., 2019; Yu et al., 2023). The availability of this infrastructure, however, does not by itself change the cognitive depth of learning. A substantial body of commentary on Chinese blended practice observes that online and offline components are often combined mechanically, with digital tasks duplicating rather than deepening classroom work, and that the resulting instruction can be lively in form while remaining shallow in substance. The concern is not that blended learning fails, but that blending without a clear pedagogical aim tends to reproduce surface learning in two media instead of one.

These two strands, the call for deeper reading and the maturation of blended infrastructure, meet in the construct of Blended Deep Learning (BDL). In earlier Chinese conceptualisations, deep learning and blended learning were treated as two constructs to be integrated on equal terms. The present study adopts a more hierarchical framing: deep learning, understood here as the cultivation of inferential, critical, and creative reading, is the pedagogical aim, and the blended structure is the delivery mechanism through which that aim is pursued across time and space. This framing has practical consequences for design, because it directs attention to what each online and offline phase is supposed to do cognitively, rather than to how much content is delivered in each medium.

Empirical evidence on blended and flipped designs for second language reading has grown quickly and is broadly encouraging. Blended instruction has been shown to improve EFL reading proficiency (Rahimzadeh & Gilakjani, 2022), flipped reading classes have yielded higher comprehension and stronger engagement (Tran et al., 2025), and meta-analytic work supports the robustness of flipped learning for L2 achievement (Aybirdi et al., 2023). Two gaps nevertheless remain. First, most of this evidence comes from secondary or tertiary learners, whose self-regulation and language resources differ markedly from those of eleven- and twelve-year-old pupils. Second, outcomes are typically reported as global comprehension or achievement scores, which cannot show whether the specific higher-order dimensions of reading that curriculum reform targets have actually changed. A study that measures inferential, critical, and creative reading separately, alongside achievement, in a primary EFL setting, addresses both gaps.

Building on the authors' earlier work, which developed and validated the Deep Reading Ability Questionnaire for primary pupils (Wen & Ahmad, 2025a) and synthesised the literature on deep learning in blended environments (Wen & Ahmad, 2025b), this study reports the effectiveness evaluation of a BDL module for Grade 6 English reading in a Chinese public primary school. The study addressed two research questions:

RQ1. What is the effect of the BDL module on primary school pupils' deep reading ability, considered as the combination of inferential, critical, and creative reading skills, and on each of these three dimensions?

RQ2. What is the effect of the BDL module on primary school pupils' English reading achievement after controlling for pre-test achievement?

Based on these research questions, the following null hypotheses were formulated. No hypothesis was formulated for the descriptive baseline comparisons.

H₀₁: There is no significant difference in the linear combination of the three dimensions of deep reading ability, which are inferential, critical, and creative reading skills, between the treatment and the control groups. (corresponding to RQ1)

H_{01a}: There is no significant difference in inferential reading skills between the treatment and the control groups.

H_{01b}: There is no significant difference in critical reading skills between the treatment and the control groups.

H_{01c}: There is no significant difference in creative reading skills between the treatment and the control groups.

H₀₂: There is no significant difference in English reading achievement between the treatment and the control groups after controlling for pre-test scores. (corresponding to RQ2).

The purpose of this study is to evaluate whether a systematically designed Blended Deep Learning (BDL) module improves primary school pupils' deep reading ability and English reading achievement, thereby demonstrating that blended infrastructure can serve as an effective mechanism for realising the thinking-quality competency in primary English education.

This study contributes to the literature in three respects. First, conceptually, it reconceptualises Blended Deep Learning as a mechanism-oriented system in which deep learning is the pedagogical aim and the three-phase blended structure the delivery mechanism, specifying the cognitive function of each phase as a corrective to mechanical media combination in Chinese blended practice. Second, empirically, it provides quasi-experimental evidence for blended EFL reading at the primary level, showing that a construct-specific outcome, namely deep reading rather than global comprehension, responds strongly to the design and thereby addressing both gaps identified above. Third, methodologically and practically, it triangulates a validated self-report measure of the three deep reading dimensions with a performance-based reading test, and, being built on existing school infrastructure and prescribed textbook topics and mapped onto the Deeper Learning Cycle, offers a replicable template for translating the thinking-quality competency into assessable classroom practice.

LITERATURE REVIEW

Deep Reading Ability as a Three-Dimensional Construct

The distinction between deep and surface approaches to learning originates in Marton and Säljö's (1976) studies of how university students processed prose passages, and it was subsequently elaborated by Biggs (1979, 1993) into a model in which deep learning combines intrinsic motivation with strategies that seek meaning, relate ideas, and check arguments. Applied to reading, this tradition converges with cognitive accounts in which comprehension is a hierarchy of increasingly demanding thinking. Hermosa (2006) describes reading comprehension as a cumulative sequence in which literal understanding is followed by interpretation, critical evaluation, application, and finally creative reading, each level requiring a higher order of thinking than the one below. In the construction-integration model, the reader builds a coherent situation model by integrating textual cues with prior knowledge, a process in which inference generation is central (Kintsch, 1988).

The present study defines deep reading ability as comprising three dimensions. Inferential reading skills refer to the ability to use stated information and context to recover implicit meaning, to predict developments, and to infer relationships such as cause and sequence. Critical reading skills refer to the ability to analyse, question, and evaluate the author's viewpoint and the text's content, distinguishing fact from opinion and forming reasoned judgements (Ocak & Karslı, 2022). Creative reading skills refer to the ability to extend, transform, or reimagine textual ideas, generating alternative solutions or endings to those presented by the author (Danesh & Nourdad, 2017). These three dimensions are directly operationalised in the Deep Reading Ability Questionnaire whose development and validation are reported in Wen and Ahmad (2025a).

Blended Deep Learning

Blended learning, at its most basic, combines face-to-face and online instruction (Graham, 2006; Garrison & Kanuka, 2004). Staker and Horn's (2012) typology of K-12 blended models distinguishes, among others, the Rotation model, in which pupils alternate between modalities within a classroom-based schedule, and the Enriched Virtual model, in which face-to-face sessions are complemented by substantial remote online learning. Research on deep learning in blended environments has grown steadily since 2017 but remains limited in volume, and it clusters around three concerns: the processes and activity designs that promote deep learning, the factors that influence it, and its evaluation (Bizami et al., 2023). A systematic review of 29 peer-reviewed studies found that deep learning in blended environments depends on interconnected conditions, including student-centred structures such as the flipped classroom, multimodal resources, strong teaching presence, scaffolded metacognitive reflection, and formative feedback, and that technology alone does not produce depth unless it is aligned with clear pedagogical goals (Wen & Ahmad, 2025b).

A useful distinction follows from this. Shallow blending redistributes conventional tasks across online and offline settings without changing their cognitive demand. Deep blending integrates the affordances of the online

component, such as asynchronous preview and multimodal scaffolding, with face-to-face cognitive work such as dialogic inference and peer-mediated evaluation, so that each phase serves a distinct step in the development of understanding. The module evaluated in this study was designed as an instance of the latter. Its three-phase sequence corresponds most closely to a combination of the Rotation and Enriched Virtual models: the in-class phase alternates teacher-led guidance with small-group collaborative activity, while the pre-class and post-class phases deliver the asynchronous virtual element through a school-adopted cloud platform.

Design Frameworks: ASSURE and the Deeper Learning Cycle

Two frameworks operating at different levels informed the module. The ASSURE model (Heinich et al., 1996) is a six-step instructional design lifecycle, comprising analyse learners, state objectives, select methods, media and materials, utilise media and materials, require learner participation, and evaluate and revise. Its emphasis on learner analysis, on the purposeful use of media, and on evaluation makes it particularly suitable for technology-integrated lesson design (Lei, 2023). The Deeper Learning Cycle (DELC) of Jensen and Nickelsen (2008) is a seven-step cognitive cycle that operates within lessons: planning standards and curriculum, pre-assessing, building a positive learning culture, priming and activating prior knowledge, acquiring new knowledge, processing the learning deeper, and evaluating learning. In the present module, ASSURE structured the development and refinement of the module as a whole, while DELC structured the cognitive progression enacted within each two-week lesson cycle. The two were aligned phase by phase: ASSURE's learner analysis corresponds to DELC pre-assessment; stating objectives corresponds to planning standards; selecting and utilising media corresponds to building a positive learning culture; requiring learner participation corresponds to priming, acquiring, and processing; and evaluate-and-revise corresponds to evaluating learning.

Theoretical Grounding

Three complementary perspectives underpin the design. Constructivism (Piaget, 1973; Vygotsky, 1978) frames learning as active meaning-making situated in social interaction, and it supplies the rationale for distributing meaning construction across preparation, collaborative interpretation, and reflection. Activity theory (Engeström, 1987) specifies how learning is mediated by an activity system of tools, rules, community, and division of labour, which in a blended module means that the platform and classroom technologies act as mediational tools, phase routines act as rules, and structured group work defines the division of labour. Comprehension as a thinking process (Hermosa, 2006; Meniado, 2016) provides the reading-specific lens that translates the general aim of deep learning into the three assessable dimensions described above. Together these perspectives inform, respectively, the module's view of learning, its structuring of mediated activity, and its specification of outcomes.

METHODOLOGY

Research Design

The study employed a quasi-experimental nonequivalent control group pretest-posttest design (Shadish et al., 2002). Random assignment of individual pupils was not feasible within normal school arrangements, so two intact Grade 6 classes were used and were randomly assigned at the class level to the treatment and control conditions. Both groups completed the Deep Reading Ability Questionnaire (DRAQ) and the English Reading Test (ERT) before and after an eight-week instructional period. The treatment group learned through the BDL module; the control group received the conventional method of instruction. To keep the comparison equitable, conventional instruction also made use of information and communication technologies such as PowerPoint slides, audio, video, and access to online search engines, so that the contrast between conditions lay in the pedagogical design rather than in the mere presence of technology.

Participants

The population comprised 490 Grade 6 pupils (233 girls and 257 boys) in a public primary school in Xuzhou City, Jiangsu Province, China. The school was selected by convenience sampling on the basis of access and feasibility, and Grade 6 classes were selected purposively because pupils at this level have foundational vocabulary and grammar and can take part consistently in structured reading tasks and blended activities.

Because the study was conducted within normal school arrangements, the available sample was determined by the two intact classes that could participate throughout the eight-week intervention. No individual-level randomisation was feasible, and an a priori power analysis was not conducted before data collection. The resulting sample therefore represents a feasibility-based classroom sample rather than a statistically powered population sample, and the limited sample size and single-school setting should be considered when interpreting the findings. The sample consisted of 60 pupils in two intact classes. The treatment group comprised 30 pupils (14 boys, 16 girls; mean age 11.45 years) and the control group 30 pupils (15 boys, 15 girls; mean age 11.50 years). Both classes were taught by the same experienced English teacher throughout the study, ensuring teacher-level equivalence across conditions. Permission was obtained from the local education authority and the school, and pupils and their guardians were informed of the purpose of the study, of the confidentiality of their data, and of their right to withdraw.

The BDL Module

The BDL module covered four reading topics mapped to the officially prescribed Grade 6 English textbook (Jiangsu edition): Cinderella, Helping Our Parents, Chinese Festivals, and Holiday Fun. Each topic was delivered through a self-contained two-week cycle. In the pre-class online phase (approximately one week), pupils previewed the text, a background video, and slides on the Xuele cloud platform using their tablets, completed a short pre-reading quiz, and could raise questions through online question-and-answer, while the teacher monitored participation remotely and used the platform data to form groups. In the 90-minute in-class phase, the teacher primed prior knowledge through brainstorming, guided silent reading with structured questions that moved from locating information to interpreting characters and motives, and then set three small-group tasks on the interactive whiteboard and tablets: locating textual evidence, retelling and performing the story from different characters' perspectives, and designing and presenting an alternative ending, with teacher scaffolding when groups' thinking stalled. In the post-class online phase (approximately one week), pupils produced a mind map organising the story structure on the platform, completed extended reading of related texts, and received remote feedback from the teacher.

Table 1 summarises how the activities of the first lesson were aligned with the three dimensions of deep reading ability, the DELC steps, and the ASSURE components; the same mapping logic governed the other three lessons. Module objectives were derived from the Compulsory Education English Curriculum Standards (2022), from the three-dimensional definition of deep reading, and from a needs analysis that drew on baseline performance, pilot feedback, and the school's technological context. Prior to the main study, three experts in English education and curriculum studies reviewed the module for alignment of objectives, tasks, and outcomes, clarity of instructions, appropriateness of difficulty and pacing, coherence between online and offline components, and transparency of assessment criteria, and the module was revised in response to their comments (Lynn, 1986). The Xuele platform, the interactive whiteboard, and the tablets were part of the school's existing infrastructure; the contribution of the module lies in the design of materials, tasks, and sequencing that exploit these affordances for deep reading.

Table 1. Alignment of Lesson 1 activities with deep reading dimensions, DELC steps, and ASSURE components

Lesson phase	Main activity	Deep reading dimension targeted	DELC step	ASSURE component
Pre-class online	Preview text, video, and slides on Xuele; complete pre-reading quiz	Inferential (activating background knowledge, predicting plot)	Pre-assessing; priming prior knowledge	Analyse learners; select and utilise media
In-class warm-up	Brainstorm fairy tales; explain preferences using "Why...? Because..."	Inferential (linking prior knowledge to the new text)	Priming prior knowledge	Require learner participation
In-class task 1	Silent reading of paragraphs 1-3; answer guided questions; group search for synonymous sentences	Inferential (identifying key information and logical relations)	Acquiring new knowledge; processing (awareness, analysis to	Require learner participation; utilise media

			synthesis)	
In-class task 2	Group role-play retelling the story from different characters' perspectives	Critical (evaluating characters' behaviour and motives); creative (retelling in own words)	Processing (analysis to synthesis; application)	Require learner participation
In-class task 3	Groups design and present a new ending	Creative (generating original endings); critical (evaluating the original ending)	Processing (application; assimilation)	Require learner participation
Post-class online	Mind map of story structure on Xuele; extended reading of related tales	Inferential (organising plot structure); creative extension	Processing the learning deeper; evaluating learning	Evaluate and revise

Instruments

Deep reading ability was measured with the Deep Reading Ability Questionnaire (DRAQ), a 30-item self-report instrument adapted from the critical and creative thinking sections of deep learning ability scale, with ten items for each of the inferential, critical, and creative reading dimensions. Pupils responded on a five-point Likert scale from strongly disagree (1) to strongly agree (5), and the instrument was administered in pupils' first language (Mandarin) to minimise language-comprehension confounds. Content validity was established by three experts (mean item-level content validity index = 0.94), face validity by ten primary school teachers, and internal consistency in a pilot study with 30 pupils not involved in the main study (Cronbach's $\alpha = 0.991, 0.990, \text{ and } 0.990$ for the three subscales; 0.997 overall). Full details of development and validation are reported in Wen and Ahmad (2025a). The unusually high alpha values reflect the deliberate conceptual homogeneity of items within each subscale, the small pilot sample, and the Likert response format, and for this reason the DRAQ was triangulated with a performance-based measure. However, the values should not be interpreted solely as evidence of superior reliability, because coefficients approaching .99 may also indicate substantial item redundancy. The DRAQ results were therefore interpreted together with its expert content validation and the performance-based ERT rather than relying on internal consistency alone. For the ERT, multiple-choice items were scored dichotomously against a key, and constructed-response items were credited when the response met the coherence criterion. The constructed-response items were evaluated using the predefined scoring criteria concerning relevance, logical coherence, and connected expression.

English reading achievement was measured with the English Reading Test (ERT) and reviewed by experts in English education (mean item-level content validity index = 0.97). Each form comprised two reading passages, each followed by three multiple-choice items and two constructed-response items, for a total of ten items scored out of 100. The five item types assessed detail comprehension, inference, main idea, evaluation of opinions, and creative expression; the last two types required pupils to write at least two coherent and logically connected sentences addressing the question. Multiple-choice items were scored dichotomously against a key, and constructed-response items were credited when the response met the coherence criterion. Pre-test and post-test forms were of comparable difficulty and were scored by the same procedure. Internal consistency in the pilot study, estimated with KR-20 (Kuder & Richardson, 1937), was 0.685, a value judged acceptable for a short test with a small pilot sample.

Procedure

The study spanned ten weeks. In the first week the teacher received a ten-hour, two-day training workshop covering blended learning, deep learning, the content of the BDL module, and its classroom use, and both groups completed the DRAQ and ERT pre-tests. The eight-week treatment followed, with the four topics taught in the same sequence to both groups, through the BDL module in the treatment group and through conventional instruction in the control group. In the final week both groups completed the DRAQ and ERT post-tests.

Data Analysis

Data were analysed with IBM SPSS Statistics Version 24 at an alpha level of .05. Independent-samples t-tests were initially used to examine baseline differences in the ERT and the three DRAQ dimensions. Because the pre-test and post-test data came from two intact classes that had been randomly assigned to conditions, the analysis of the pretest-posttest data proceeded in two stages. First, baseline equivalence on every pre-test measure was verified with independent-samples t-tests supplemented by standardised mean differences (Cohen's *d*) with 95% confidence intervals, and the homogeneity of the pre-test to post-test relationship across groups was checked for each deep reading dimension. Second, given the verified equivalence, a one-way MANOVA was conducted on the three post-test DRAQ dimensions with group as the fixed factor, followed by univariate ANOVAs for individual dimensions when the multivariate group effect was significant. This approach examined the combined effect of the intervention across the three related dimensions while controlling the family-wise error rate.

For English reading achievement, a one-way ANCOVA was conducted on post-test ERT scores with group as the fixed factor and pre-test ERT score as the covariate, which adjusts for initial differences and yields a more precise estimate of the treatment effect. Partial eta squared was reported as the effect size for each model and interpreted against the conventional benchmarks of .01 (small), .06 (medium), and .14 (large) (Cohen, 1988; Pallant, 2020). Post-test group differences were additionally expressed as Cohen's *d* with 95% confidence intervals so that the magnitude and precision of each effect could be evaluated independently of significance testing. For the MANOVA, assumptions of multivariate normality, linearity, and homogeneity of variance-covariance matrices were examined before the analysis.

Ethical Considerations

The study was conducted in accordance with the ethical principles of voluntary participation, informed consent, confidentiality, and respect for participants. The study protocol, including the participant information sheets, consent forms, and all instruments, was reviewed and approved by the Human Research Ethics Committee of Universiti Sains Malaysia (Jawatankuasa Etika Penyelidikan Manusia, JEPeM-USM; approval code USM/JEPeM/PP/24060541). Permission to conduct the study was also obtained from the relevant school authorities before data collection. The purpose and procedures of the study were explained to the participating pupils and their guardians, and participation was voluntary. Participants were informed that they could withdraw from the study without penalty. No personally identifying information was collected, and all questionnaire and assessment data were treated confidentially and used solely for research purposes.

RESULTS

Assumption Checks and Baseline Equivalence

Shapiro-Wilk tests indicated no significant departures from normality for the three DRAQ dimensions in either group at pre-test (treatment: $p = .074, .138, .087$; control: $p = .424, .274, .902$ for inferential, critical, and creative reading respectively) or for ERT pre-test scores (treatment $p = .940$; control $p = .855$). Linearity between each pre-test and corresponding post-test dimension was confirmed by scatterplots, and the homogeneity of regression slopes across groups was supported for every dimension, as the Group $\times$ pre-test interaction was non-significant for inferential, $F(1, 56) = 0.006, p = .936$, critical, $F(1, 56) = 0.048, p = .827$, and creative reading, $F(1, 56) = 0.683, p = .412$, indicating that the relationship between pre-test and post-test scores did not differ between the two groups. Post-test correlations among the three dimensions were small and non-significant in both groups (largest $|r| = .30$), indicating no multicollinearity. Boxplots identified two mild univariate outliers that were retained because they were not extreme and exerted no undue influence. Box's *M* test supported the homogeneity of variance-covariance matrices, $M = 0.637, p = .996$, and Levene's tests for the three post-test dimensions were all non-significant ($p = .950, .979$, and $.836$). Linearity of the covariate-outcome relationship and homogeneity of regression slopes were examined and found to be satisfactory before the ANCOVA was conducted.

Table 2 presents the baseline comparisons. The two groups did not differ significantly at pre-test on inferential reading, $t(58) = 0.05, p = .963$, critical reading, $t(58) = 0.31, p = .756$, creative reading, $t(58) = 1.60, p = .116$,

or ERT achievement, $t(58) = -1.28$, $p = .205$. The groups therefore entered the study with comparable levels of deep reading ability and reading achievement. Standardised mean differences at pre-test were negligible for inferential reading ($d = 0.03$, 95% CI [-0.48, 0.53]) and critical reading ($d = 0.08$, 95% CI [-0.42, 0.59]), and small to moderate but non-significant for creative reading ($d = 0.41$, 95% CI [-0.10, 0.92]) and ERT achievement ($d = -0.33$, 95% CI [-0.84, 0.18]), with every confidence interval including zero. Together with the homogeneity of regression slopes reported above, these results support the interpretation of post-test group differences on the deep reading dimensions as effects of the intervention rather than of pre-existing differences.

Table 2. Pre-test comparisons between the treatment and control groups

Measure	Group	M	SD	t(58)	p
Inferential reading skills	Treatment	3.20	0.35	0.05	.963
	Control	3.19	0.41		
Critical reading skills	Treatment	3.14	0.35	0.31	.756
	Control	3.11	0.38		
Creative reading skills	Treatment	3.11	0.47	1.60	.116
	Control	2.92	0.46		
English Reading Test	Treatment	67.58	4.89	-1.28	.205
	Control	69.21	4.98		

Descriptive Statistics

Table 3 reports the means and standard deviations for all measures before and after the intervention. At post-test, the treatment group's overall deep reading ability rose from 3.15 to 3.74, whereas the control group's changed from 3.07 to 3.13. The pattern was consistent across the three dimensions: inferential reading rose from 3.20 to 4.09 in the treatment group against 3.19 to 3.30 in the control group; critical reading from 3.14 to 3.70 against 3.11 to 3.19; and creative reading from 3.11 to 3.43 against 2.92 to 2.90. On the ERT, the treatment group improved by 10.63 points (67.58 to 78.21) compared with 1.38 points in the control group (69.21 to 70.59). Standard deviations remained similar across groups and occasions, indicating that the treatment group's gains were broadly shared rather than driven by a few pupils.

Table 3. Means and standard deviations at pre-test and post-test by group

Measure	Group	Pre-test M (SD)	Post-test M (SD)
Overall deep reading ability	Treatment	3.15 (0.23)	3.74 (0.25)
	Control	3.07 (0.23)	3.13 (0.24)
Inferential reading skills	Treatment	3.20 (0.35)	4.09 (0.46)
	Control	3.19 (0.41)	3.30 (0.43)
Critical reading skills	Treatment	3.14 (0.35)	3.70 (0.43)
	Control	3.11 (0.38)	3.19 (0.43)
Creative reading skills	Treatment	3.11 (0.47)	3.43 (0.58)
	Control	2.92 (0.46)	2.90 (0.51)
English Reading Test (out of 100)	Treatment	67.58 (4.89)	78.21 (5.29)
	Control	69.21 (4.98)	70.59 (5.20)

Note. DRAQ scores are item means on a five-point scale ($n = 30$ per group).

Effect on Deep Reading Ability (RQ1)

A one-way MANOVA with group as the independent variable and the three post-test DRAQ dimensions as dependent variables revealed a statistically significant multivariate effect of the BDL module, Wilks' $\Lambda = 0.3707$, $F(3, 56) = 31.69$, $p < .001$, partial $\eta^2 = 0.6293$ (Table 4). Pillai's trace (0.6293), Hotelling's trace (1.6975), and Roy's largest root (1.6975) led to the same conclusion, all significant at $p < .001$. The effect is large by conventional standards, indicating that group membership accounted for about 63% of the variance in the linear combination of the three dimensions. Null hypothesis H_{01} was therefore rejected.

Table 4. Multivariate tests for post-test deep reading ability by group

Test statistic	Value	F	Hypothesis df	Error df	p	Partial η^2
Pillai's trace	0.6293	31.687	3	56	< .001	0.6293
Wilks' lambda	0.3707	31.687	3	56	< .001	0.6293
Hotelling's trace	1.6975	31.687	3	56	< .001	0.6293
Roy's largest root	1.6975	31.687	3	56	< .001	0.6293

Follow-up univariate ANOVAs (Table 5) showed that the treatment group outperformed the control group on every dimension. For inferential reading skills the effect was $F(1, 58) = 46.76$, $p < .001$, partial $\eta^2 = 0.446$, so that instructional condition explained approximately 45% of post-test variance. For critical reading skills the effect was $F(1, 58) = 21.73$, $p < .001$, partial $\eta^2 = 0.273$, and for creative reading skills $F(1, 58) = 13.78$, $p < .001$, partial $\eta^2 = 0.192$. Expressed as standardised mean differences, the post-test group differences were $d = 1.77$, 95% CI [1.18, 2.37] for inferential reading, $d = 1.19$, 95% CI [0.64, 1.73] for critical reading, and $d = 0.97$, 95% CI [0.44, 1.51] for creative reading, with no interval approaching zero. All three effects exceed the threshold for a large effect, although their magnitude decreased from inferential to critical to creative reading. Null hypotheses H_{01a} , H_{01b} , and H_{01c} were rejected.

Table 5. Univariate ANOVAs for the three post-test deep reading dimensions

Dependent variable	Source	SS	df	MS	F	p	Partial η^2
Inferential reading skills	Group	9.394	1	9.394	46.76	< .001	0.446
	Error	11.653	58	0.201			
Critical reading skills	Group	3.989	1	3.989	21.73	< .001	0.273
	Error	10.647	58	0.184			
Creative reading skills	Group	4.112	1	4.112	13.78	< .001	0.192
	Error	17.313	58	0.299			

Effect on English Reading Achievement (RQ2)

A one-way ANCOVA on post-test ERT scores, with pre-test ERT score as the covariate, revealed a statistically significant main effect of group after adjusting for baseline achievement, $F(1, 57) = 225.76$, $p < .001$, partial $\eta^2 = 0.799$ (Table 6). The unadjusted post-test difference corresponded to $d = 1.45$, 95% CI [0.88, 2.02]. The covariate was also significant, $F(1, 57) = 235.77$, $p < .001$, confirming that pre-test achievement predicted post-test achievement and that its statistical control was warranted. Evaluated at the grand mean of the covariate (68.39), the adjusted post-test mean of the treatment group was 78.99 compared with 69.81 for the control group, a difference of about nine points on a 100-point test. Null hypothesis H_{02} was rejected.

Table 6. ANCOVA for post-test English reading achievement with pre-test score as covariate

Source	SS	df	MS	F	p	Partial η^2
Pre-test ERT (covariate)	1284.09	1	1284.09	235.77	< .001	
Group	1229.60	1	1229.60	225.76	< .001	0.799
Error	310.44	57	5.45			

Note. Adjusted post-test means: treatment = 78.99, control = 69.81 (covariate evaluated at 68.39).

DISCUSSION

Deep Reading Ability

The multivariate result indicates that, after controlling for baseline deep reading ability, pupils who received the BDL module showed substantially higher post-test deep reading outcomes than those who received conventional instruction. Given the quasi-experimental design, this difference should be interpreted as evidence consistent with an intervention effect rather than as definitive proof of causality. Interpreted through the deep-surface

distinction, this outcome is plausible because the module was structured to make active meaning construction unavoidable rather than optional (Marton & Säljö, 1976; Biggs, 1993). Pupils could not complete the cycle by memorising and reproducing; each phase required them to prepare, interact, and reflect on the text in ways that align with the deep approach. The direction of the finding converges with recent evidence that blended and flipped designs improve L2 reading outcomes when pre-class input is combined with in-class active processing (Rahimzadeh & Gilakjani, 2022; Tran et al., 2025; Aybirdi et al., 2023). What the present study adds is evidence that a construct-specific outcome, deep reading rather than global comprehension, responds strongly to such a design in a primary EFL population, where self-regulation and linguistic resources are more limited than in the adolescent and adult samples that dominate the literature.

The ordering of the three univariate effects deserves comment. The largest gain was in inferential reading, followed by critical and then creative reading. This gradient mirrors the cumulative hierarchy of comprehension in which each level builds on the one below (Hermosa, 2006), and it is consistent with the allocation of instructional time in the module. Inferential work was scaffolded in every phase: schema activation and prediction before class, evidence-based questioning during class, and structural organisation through mind maps afterwards. This pattern resembles the conditions under which Samiei and Ebadi (2021) observed gains in inferential comprehension through a flipped design, where relocating initial input to the pre-class phase freed classroom time for reasoning, and it is consistent with the finding that explicit inferential strategy instruction delivered online improves inferential comprehension when it is clearly organised and accompanied by feedback (Sanhueza-Oyarzún & Quintanilla-Espinoza, 2024). Within the construction-integration account, such scaffolding directly supports the integration of textual cues with prior knowledge on which inference depends (Kintsch, 1988).

Critical reading was cultivated mainly in the in-class phase through role-play from different characters' perspectives, evaluation of characters' motives, and dialogic questioning, which positions comprehension as an evaluative and social process (Vygotsky, 1978). The gain is in line with reports that flipped and web-based designs improve critical thinking for critical reading when class time is used for evaluative discussion rather than content delivery (Yulian, 2021; Pang, 2022), and with evidence that multimodal tasks can prompt learners to interrogate perspectives and assumptions (Lam & Putri, 2024). The smaller, though still large, effect for creative reading is understandable given that creative response was concentrated in a single in-class task and a post-class extension, and given that generative transformation of text is the most demanding level of the hierarchy. Even so, the result accords with recent findings that creative reading practices and digital storytelling enhance higher-order engagement with texts (Shahini et al., 2025; Hidayat et al., 2024) and that role-based interaction and sustained peer dialogue cultivate creative textual responses (Baki, 2025).

Three qualifications apply to the interpretation of all three effects. The participants were primary pupils, so the findings should not be assumed to transfer to older learners, or vice versa, without further testing (Samiei & Ebadi, 2021; Song et al., 2025). The module was a multi-component intervention, and the gains are best attributed to the combined influence of instructional sequencing, task design, and teacher mediation rather than to any single activity or digital feature, a point that meta-analytic work on flipped and technology-based interventions has repeatedly made (Strelan et al., 2020; Alqahtani, 2024). Finally, the DRAQ is closely aligned with the objectives embedded in the module, and strong instruction-assessment correspondence can inflate observed effects relative to more distal measures (AlManafi et al., 2023). The findings are therefore evidence for the effectiveness of a structured BDL pathway under specific learner and implementation conditions, not proof that blending in itself enhances deep reading.

English Reading Achievement

After adjustment for pre-test achievement, the treatment group scored about nine points higher on the ERT than the control group, with a large effect size. Because the ERT is a performance measure rather than a self-report, this result provides converging evidence that the changes captured by the DRAQ were accompanied by measurable improvement in what pupils could actually do with a text. The finding is consistent with evidence that blended formats improve reading comprehension when multimedia pre-reading support and interactive in-class processing are emphasised (Bersamin et al., 2024), and with the effect of blended learning on primary-stage EFL reading achievement reported by AlManafi et al. (2023). It is also consistent with the module's use of

immediate, task-focused feedback through the platform and in group tasks, a feature whose value depends on feedback being specific and tied to the learning task rather than generic (Kluger & DeNisi, 1996; Hattie et al., 2016). Two features of the design strengthen the causal interpretation: the same teacher taught both groups, and the control group also used technology, so the observed difference may be related to the pedagogical integration of the module rather than to technology exposure alone, although selection, teacher expectancy, and contextual effects cannot be completely excluded.

Implications

For theory, the study supports conceptualising BDL not as an additive combination of two constructs but as a mechanism-oriented learning system in which phase sequencing provides temporal scaffolding for constructivist meaning-making, the mediated activity system of tools, rules, and division of labour is deliberately reconfigured (Engeström, 1987), and the hierarchy of comprehension processes specifies what "deep" means in reading (Hermosa, 2006; Kintsch, 1988). The differentiated pattern of effects across the three dimensions illustrates why outcome measures in blended reading research should be construct-specific; a single achievement score would have concealed the gradient reported here.

For practice, the results suggest that the decisive factor is instructional integration rather than technology adoption. Pre-class tasks should be short and purposeful, and they work best when pupils arrive in class with a prediction, a question, or a piece of textual evidence. Classroom time should be reserved for interaction that requires evidence-based interpretation, perspective-taking, and evaluation, with teacher questioning that systematically moves beyond literal recall. Post-class tasks should require reorganisation of meaning, such as mind maps or alternative endings, rather than repetition. Because the module was built on infrastructure the school already possessed and on topics from the prescribed textbook, it offers a replicable template for teachers working under similar curricular and technological conditions, and the ASSURE and DELC alignment provides a transparent procedure for extending the design to new units.

Limitations and Future Research

Several limitations bound these conclusions. First, the study involved only two intact classes from a single public primary school, with 30 pupils in each group. Although the classes were randomly assigned to treatment and control conditions and baseline differences were not statistically significant, the small, single-school sample limits the generalisability of the findings. Second, an a priori power analysis was not conducted, and the sample was determined primarily by the available intact classes and practical feasibility. Future studies should use a larger number of classes and schools and conduct an a priori power analysis before data collection. Third, although baseline equivalence was established and the achievement analysis controlled for pre-test scores, the quasi-experimental design cannot completely eliminate selection, teacher expectancy, or contextual effects. The same teacher taught both conditions, which reduced teacher-level variation but also created a potential risk of expectancy or treatment contamination. No formal independent treatment-fidelity measure was collected, so implementation fidelity could not be quantified. In addition, the three deep reading dimensions were compared at post-test on the basis of verified baseline equivalence and homogeneous regression slopes rather than through covariate adjustment; although the pre-test standardised differences were small and non-significant, this approach does not remove residual baseline variance from the estimate of the intervention effect, and future studies with larger samples should model pre-test scores directly. Fourth, the DRAQ showed extremely high internal consistency, which may partly reflect item redundancy, while the ERT showed modest internal consistency ($KR-20 = 0.685$). The constructed-response component would benefit from independent double scoring and formal inter-rater reliability assessment in future research. Finally, the intervention covered four topics over eight weeks, and only immediate post-tests were administered. Future research should replicate the module across schools and regions, extend the intervention across grades and reading genres, include delayed post-tests, incorporate performance-based and process-sensitive measures, and collect independent fidelity observations and classroom-process data.

CONCLUSION

This study evaluated a Blended Deep Learning module for Grade 6 English reading in a Chinese public primary

school, in which deep learning was treated as the pedagogical aim and a three-phase blended structure as the delivery mechanism. Compared with conventional instruction that also used technology, the module produced large and statistically significant improvements in pupils' deep reading ability, in each of its inferential, critical, and creative dimensions, and in English reading achievement after adjustment for pre-test scores. The gradient of effects across the three dimensions was consistent with the hierarchy of comprehension and with the distribution of instructional emphasis across the module's phases. Within the limits of a single-site quasi-experiment, the findings indicate that blended learning cultivates higher-order reading in primary EFL classrooms when it is implemented as an integrated instructional system with clear phase routines, deliberately structured classroom discourse, and consolidation tasks that require pupils to reconstruct meaning. The module, its alignment with established design frameworks, and the construct-specific measurement approach reported here provide a foundation for further research and for classroom practice aimed at moving young readers beyond the surface of the text.

REFERENCES

1. AlManafi, A. O. S., Osman, S. Z. M., & Magableh, I. S. I. (2023). The effect of blended learning on the primary stage EFL students' reading comprehension achievement in Libya. *International Journal of Instruction*, 16(2), 703-718.
2. Alqahtani, S. S. (2024). A meta-analysis of technology-based interventions for elementary students with reading difficulties. *Humanities and Social Sciences Communications*, 11(1), 1-10.
3. Aybirdi, N., Efe, H., & Atasoy Sal, Ç. (2023). The impact of flipped learning on L2 learners' achievements: A meta-analysis. *International Journal of Education*, 11, 41-60.
4. Baki, Y. (2025). The effect of the reading circle method on curiosity and exploration, creative reading and visual literacy. *Journal of Intelligence*, 13(7), 74.
5. Bersamin, A. E., Ulla, M. B., Saripa, A., Suebsom, K., & Lemana II, H. E. (2024). Blended learning and its impact on English reading comprehension among Thai vocational students. *rEFLections*, 31(1), 165-185.
6. Biggs, J. (1979). Individual differences in study processes and the quality of learning outcomes. *Higher Education*, 8(4), 381-394.
7. Biggs, J. (1993). What do inventories of students' learning processes really measure? A theoretical review and clarification. *British Journal of Educational Psychology*, 63(1), 3-19.
8. Bizami, N. A., Tasir, Z., & Kew, S. N. (2023). Innovative pedagogical principles and technological tools capabilities for immersive blended learning: A systematic literature review. *Education and Information Technologies*, 28(2), 1373-1425.
9. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
10. Engeström, Y. (1987). *Learning by expanding: An activity-theoretical approach to developmental research*. Orienta-Konsultit.
11. Field, A. (2024). *Discovering statistics using IBM SPSS Statistics* (6th ed.). Sage.
12. Garrison, D. R., & Kanuka, H. (2004). Blended learning: Uncovering its transformative potential in higher education. *The Internet and Higher Education*, 7(2), 95-105.
13. Geng, S., Law, K. M., & Niu, B. (2019). Investigating self-directed learning and technology readiness in blending learning environment. *International Journal of Educational Technology in Higher Education*, 16(1), 1-22.
14. Graham, C. R. (2006). Blended learning systems: Definition, current trends, and future directions. In C. J. Bonk & C. R. Graham (Eds.), *The handbook of blended learning: Global perspectives, local designs* (pp. 3-21). Pfeiffer.
15. Hattie, J., Fisher, D., Frey, N., Gojak, L. M., Moore, S. D., & Mellman, W. (2016). *Visible learning for mathematics, grades K-12: What works best to optimize student learning*. Corwin Press.
16. Heinich, R., Molenda, M., Russell, J. D., & Smaldino, S. E. (1996). *Instructional media and technologies for learning* (5th ed.). Prentice Hall.
17. Hermosa, N. N. (2006). *The psychology of reading*. University of the Philippines Open University.

18. Hidayat, N. H., Sundari, H., & Zuhairi, A. (2024). EFL reading comprehension through digital storytelling in multimodal digital literacy. *Scope: Journal of English Language Teaching*, 9(1), 380-391.
19. Jensen, E., & Nickelsen, L. (2008). *Deeper learning: 7 powerful strategies for in-depth and longer-lasting learning*. Corwin Press.
20. Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2), 163-182.
21. Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254-284.
22. Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2(3), 151-160.
23. Lam, K. Y., & Putri, E. W. (2024). A study on the affordances of digital fairy-tale rewriting: A critical literacy approach to digital multimodal composing. *System*, 127, 103536.
24. LaRusso, M., Kim, H. Y., Selman, R., Uccelli, P., Dawson, T., Jones, S., Donovan, S., & Snow, C. (2016). Contributions of academic language, perspective taking, and complex reasoning to deep reading comprehension. *Journal of Research on Educational Effectiveness*, 9(2), 201-222.
25. Lei, G. (2023). Influence of ASSURE model in enhancing educational technology. *Interactive Learning Environments*, 1-17.
26. Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, 35(6), 382-386.
27. Marton, F., & Säljö, R. (1976). On qualitative differences in learning: I. Outcome and process. *British Journal of Educational Psychology*, 46(1), 4-11.
28. Meniado, J. C. (2016). Metacognitive reading strategies, motivation, and reading comprehension performance of Saudi EFL students. *English Language Teaching*, 9(3), 117-129.
29. Ocak, G., & Karşlı, E. (2022). Relationship between critical reading skills and creative reading perceptions of fifth grade students. *International Journal of Education and Literacy Studies*, 10(3), 91-100.
30. Pallant, J. (2020). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS (7th ed.)*. Routledge.
31. Pang, Y. (2022). The role of web-based flipped learning in EFL learners' critical thinking and learner engagement. *Frontiers in Psychology*, 13, 1008257.
32. Piaget, J. (1973). *To understand is to invent: The future of education*. Grossman.
33. Rahimzadeh, L., & Gilakjani, A. P. (2022). Investigating the effectiveness of blended learning in reading proficiency: Evidence from intermediate EFL learners. *Knowledge Management & E-Learning*, 14(4), 489-509.
34. Samiei, F., & Ebadi, S. (2021). Exploring EFL learners' inferential reading comprehension skills through a flipped classroom. *Research and Practice in Technology Enhanced Learning*, 16(1), 12.
35. Sanhueza-Oyarzún, L. A., & Quintanilla-Espinoza, A. (2024). The contribution of inferential reading strategies to improve the reading comprehension skill in an online private lesson context. *Revista Tecnológica-Educativa Docentes 2.0*, 17(2), 159-169.
36. Shadish, W. R., Cook, T. D., & Campbell, T. D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.
37. Shahini, G., Toroujeni, S. M. H., & Ebrahimi, M. R. (2025). Creative reading's impact on EFL learners' critical thinking and reading comprehension. *Thinking Skills and Creativity*, 101855.
38. Song, X., Razali, A. B., Sulaiman, T., & Jeyaraj, J. J. (2025). Effectiveness of online project-based learning on Chinese EFL learners' critical thinking skills and reading comprehension ability. *Thinking Skills and Creativity*, 56, 101778.
39. Staker, H., & Horn, M. B. (2012). *Classifying K-12 blended learning*. Innosight Institute.
40. Strelan, P., Osborn, A., & Palmer, E. (2020). The flipped classroom: A meta-analysis of effects on student performance across disciplines and education levels. *Educational Research Review*, 30, 100314.
41. Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics (7th ed.)*. Pearson.
42. Tran, Q., Verezub, E., & Fisher, R. (2025). Flipping EFL reading comprehension classes: Students' learning achievement and perceptions. *Teaching and Learning Inquiry*, 13, 1-44.

43. Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.
44. Wen, Y., & Ahmad, N. J. (2025a). Development, validity and reliability of Deep Reading Ability Questionnaire for primary school pupils. *International Journal of Modern Education*, 7(24), 933-949. <https://doi.org/10.35631/IJMOE.724067>
45. Wen, Y., & Ahmad, N. J. (2025b). Deep learning in blended learning environments: A systematic literature review. *Global Journal of Educational Research and Management*, 5(2), 1-22.
46. Wolf, M., & Barzillai, M. (2009). The importance of deep reading. *Educational Leadership*, 66(6), 32-37.
47. Yu, T., Dai, J., & Wang, C. (2023). Adoption of blended learning: Chinese university students' perspectives. *Humanities and Social Sciences Communications*, 10(1), 1-16.
48. Yulian, R. (2021). The flipped classroom: Improving critical thinking for critical reading of EFL learners in higher education. *Studies in English Language and Education*, 8(2), 508-522.