EDUTECH

Jurnal Teknologi Pendidikan

Journal homepage https://ejournal.upi.edu/index.php/edutech

Designing Inference Instruction for Reading Comprehension in Grades 3–6: A Rapid Systematic Review of Intervention Strategies, Constructive Alignment, and Research Gaps

Agrita Putri Nadia[1*], Tri Suminar[1], Budiyono[1], Rafika Bayu Kusumandari[1]

[1]Department of Curriculum Development, Faculty of Education and Psychology, Universitas Negeri Semarang, Semarang, Indonesia

*Correspondence: e-mail: agritaputrinadia@students.unnes.ac.id

ABSTRACT

Inference-making is central to reading comprehension, yet literacy instruction in elementary schools remains largely oriented toward locating explicitly stated information. This rapid systematic review mapped the intervention designs used to teach inference in Grades 3–6, examined the degree of alignment among their instructional components, and identified gaps for future research. The review was reported in accordance with the PRISMA 2020 statement and used streamlined, rapid-review procedures: records were identified through Google Scholar and backward citation chasing, and screening and coding were performed by a single reviewer. Of the 36 records assessed for eligibility, seven studies met the core inclusion criteria, and two early-grade studies were examined in a supplementary analysis. Data were coded against a framework comprising six instructional components derived from curriculum development and constructive alignment theory and five inferential thinking indicators derived from the literature on inference in reading, and methodological quality was appraised with the Mixed Methods Appraisal Tool. Four patterns emerged. First, international studies targeted specific inference types, whereas Indonesian studies tended to treat inference as one facet of general comprehension. Second, the most consistent evidence, judged to be of moderate certainty, came from explicit inference instruction delivered through inferential questioning, graphic organizers, and explanatory feedback; close reading showed no clear advantage. Third, few studies explicitly designed all six instructional components, with assessment the component most often left implicit. Fourth, digital learning materials remained largely untested for inferential outcomes in the upper elementary grades. Because

ARTICLE INFO

Article History:
Submitted/Received 7 Sept 2026
First Revised 19 Sept 2026
Accepted 1 Oct 2026
First Available online 31 Oct 2026
Publication Date 31 Oct 2026

Keyword:
Constructive alignment,
Elementary education,
Inference instruction,
Rapid systematic review,
Reading comprehension.

the search relied on a single search engine and only two of the seven studies were of high quality, these patterns should be regarded as tentative. The available evidence suggests that instructional designs may benefit from combining explicit inference instruction with close alignment among learning objectives, activities, materials, and assessment.
© 2025 Teknologi Pendidikan UPI

1. INTRODUCTION

In international assessment frameworks, reading literacy is no longer equated with the retrieval of explicitly stated information. The PISA 2022 framework positions understanding, integrating, evaluating, and reflecting on texts as the core of reading literacy (OECD, 2023a). Grasping implicit meaning requires readers to generate inferences, that is, to connect information in the text with other textual cues and with prior knowledge in order to construct a coherent mental representation (Kintsch, 1998). PISA 2022 results indicate that most Indonesian students perform below the minimum proficiency level in reading (OECD, 2023b), suggesting that the capacity to integrate information and draw conclusions is not adequately developed during the elementary years.

Meta-analytic evidence shows that inference-making can be taught. Elleman (2017) reported positive effects of inference instruction on both general and inferential comprehension. In classroom settings, however, effectiveness depends not only on the choice of strategy but also on the alignment among learning objectives, activities, materials, and assessment (Biggs, 1996; Wiggins & McTighe, 2005). An intervention that trains one inference type while measuring general comprehension, for instance, may fail to capture its actual impact.

Several syntheses have examined inference instruction. Hall (2016) synthesized nine intervention studies with struggling readers and reported large effects on researcher-developed inferential measures but more variable effects on standardized comprehension tests. Elleman (2017) meta-analyzed inference instruction and examined whether its effects on literal, inferential, and general comprehension differed between skilled and less skilled readers. Broader syntheses of reading interventions for struggling readers in the upper elementary grades (Wanzek et al., 2010) and of comprehension strategy instruction (Peng et al., 2023) likewise focused on effect sizes, reader characteristics, and combinations of strategies, and a recent review of close reading in Grades 3–9 judged its evidence base to be still inconclusive (Bogaerds-Hazenberg et al., 2026). These reviews share two features. First, they treat interventions as packages of strategies and ask whether, and for whom, they work; to our knowledge, none has coded the extent to which an intervention's objectives, content, activities, materials, and assessment are designed in alignment with one another. Second, they drew on studies conducted mainly in North American and European contexts and did not include research conducted in Indonesia. Yet the instructional design perspective is precisely what curriculum developers and teachers need when translating research evidence into implementable instructional designs, and the gap between proximal and standardized outcomes noted by Hall (2016) suggests that the alignment between what is trained and what is measured may partly explain variation in reported effects.

1.1. Conceptual Framework

Constructive alignment, as originally formulated by Biggs (1996), links three elements: intended learning outcomes, teaching and learning activities, and assessment

DOI: https://doi.org/10.17509/e.v25i3.106824
P-ISSN 0852-1190 E-ISSN 2502-0781

tasks. Because the present review analyzed classroom interventions rather than university courses, this three-element core was expanded using curriculum development theory. Tyler (1949) framed curriculum design as the coordination of objectives, learning experiences, their organization, and evaluation; Taba (1962) added the explicit selection and organization of content and grounded these decisions in a diagnosis of learners' needs; and backward design (Wiggins & McTighe, 2005) requires that assessment evidence and learning activities be derived from the intended outcomes. Combining these perspectives yields six instructional components: (1) learning objectives, (2) content, (3) instructional model or strategy, (4) learning activities, (5) instructional materials, and (6) assessment. These components also correspond to the elements that Indonesian lesson plans are required to specify under the national process standard (Kementerian Pendidikan dan Kebudayaan, 2016), which makes the framework directly usable by Indonesian curriculum developers and teachers. An intervention was regarded as fully aligned when all six components were explicitly designed to target the same inferential process.

The five inferential thinking indicators were derived from Kispal's (2008) review of inference in reading, which defines inference as linking textual information with prior knowledge to reach conclusions that are not explicitly stated, and from established distinctions among inference types. *Implicit meaning* corresponds to gap-filling or elaborative inferences that draw on background knowledge (Cain & Oakhill, 1999; Kispal, 2008); *connecting information* corresponds to text-connecting or bridging inferences that integrate propositions across a text (Cain & Oakhill, 1999; Graesser et al., 1994); *cause and effect* corresponds to causal inferences that maintain text coherence (Graesser et al., 1994; Kintsch, 1998); *prediction* corresponds to predictive or forward inferences (Graesser et al., 1994); and *evidence-based conclusions* correspond to the expectation that readers justify their inferences with textual evidence (Kispal, 2008; National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010). The operational definitions used for coding are given in the Methods.

Accordingly, this review addressed four questions:

RQ1: How is inferential ability conceptualized and measured in literacy interventions for Grades 3–6?

RQ2: Which instructional models, strategies, and materials are employed, and what do the findings indicate about their effectiveness?

RQ3: To what extent do these interventions align the six instructional components defined in the conceptual framework?

RQ4: Which research gaps should future studies address?

2. METHODS

2.1. Design and Reporting

This study was a rapid systematic review with thematic synthesis. Reporting followed the PRISMA 2020 statement (Page et al., 2021); the completed PRISMA 2020 checklist is provided as Supplementary Material 3. Because the review was conducted within the time and resource constraints of a master's thesis project, it used streamlined procedures of the kind described for rapid reviews (Garritty et al., 2021): a single search source and a single reviewer. These deviations from a full systematic review are reported transparently below and in the Limitations. No formal review protocol was prepared or registered.

DOI: https://doi.org/10.17509/e.v25i3.106824
P-ISSN 0852-1190 E-ISSN 2502-0781

2.2. Review Framework and Eligibility Criteria

The scope of the review was defined using the PICOC framework. The *population* comprised students in Grades 3–6. This range follows the convention of treating Grades 3–4 as the transition from learning to read to reading to learn (Chall, 1983) and matches the grade bands commonly used in upper elementary intervention research. Because the "upper grades" (*kelas tinggi*) of Indonesian elementary schools conventionally refer to Grades 4–6, the only Grade 3 study (S5) was flagged, and the robustness of the conclusions to its removal was examined in the sensitivity analysis. *Interventions* were literacy instruction explicitly targeting inferential or analytical processes. *Comparisons* included business-as-usual instruction, alternative interventions, or no comparison. Eligible *outcomes* were inferential ability, critical thinking in reading, or general reading comprehension where the intervention explicitly targeted inferential or analytical processes. The *context* was elementary education, including digital learning environments.

Table 1. Eligibility criteria.

Code	Inclusion	Code	Exclusion
I-1	Peer-reviewed journal article	E-1	Non-peer-reviewed proceedings, undergraduate or graduate theses
I-2	Published 2012–2026	E-2	Published before 2012
I-3	Primary empirical intervention study (experimental, quasi-experimental, or single-case)	E-3	Secondary review, practitioner article, or non-intervention study
I-4	Students in Grades 3–6	E-4	Secondary-school students, adult or EFL learners
I-5	Outcome consistent with the PICOC framework	E-5	Intervention or outcome not targeting inferential or analytical processes
I-6	Written in English or Indonesian	E-6	Full text, bibliographic metadata, or reported data could not be verified
		E-7	Classroom action research without a comparison group

The lower bound of 2012 was chosen because it follows the curriculum reforms of the early 2010s that made inferential and text-dependent comprehension explicit instructional targets in the United States (National Governors Association Center for Best Practices & Council of Chief State School Officers, 2010) and in Indonesia (Kurikulum 2013; Kementerian Pendidikan dan Kebudayaan, 2016), so that interventions published from 2012 onward were designed under comparable curricular expectations. Studies of Kindergarten to Grade 2 that otherwise met the criteria were analyzed separately in a supplementary contextual analysis; they did not contribute to the core synthesis.

2.3. Information Sources and Search Strategy

Searches were conducted on 14 October 2026 in Google Scholar using nine substantive queries in English and Indonesian. The queries combined terms such as *inference instruction*, *inference-making*, *close reading*, *problem-based learning*, *critical reading*, *computer-based inference training*, "pemahaman inferensial" (inferential comprehension), "e-modul" (e-module), and "berpikir kritis" (critical thinking) with

DOI: https://doi.org/10.17509/e.v25i3.106824
P-ISSN 0852-1190 E-ISSN 2502-0781

descriptors of elementary schooling. Terms for problem-based learning and e-modules were included under the Intervention and Context elements of the PICOC framework, respectively, to capture instructional models that explicitly target analytical thinking and the digital materials commonly used in Indonesian elementary schools. These searches were supplemented by seven metadata verification queries and by backward citation chasing from the secondary reviews identified. The verification queries were used to confirm the bibliographic details of studies already identified; because they targeted known records, any records they returned were duplicates or non-article entries. Figure 1 therefore reports all 16 queries at the identification stage. Bibliographic databases (e.g., Scopus, ERIC, Web of Science, SINTA, and Garuda) were not searched; the implications of this decision are discussed under Limitations. The full query list is provided in Supplementary Material 1.

2.4. Study Selection

The searches yielded 147 records. Duplicates and non-article entries (e.g., blogs, podcasts, product pages, and research-grant pages) were removed (n = 111); these two categories were not tallied separately. The high proportion of non-article entries reflects the use of a general academic search engine rather than a curated bibliographic database. Screening, eligibility assessment, data extraction, and quality appraisal were all performed by the first author; no second reviewer was involved, so inter-rater agreement could not be estimated. The remaining 36 records were screened and assessed for eligibility against the criteria in Table 1 in a single step, on the basis of their abstracts and, where accessible, their full texts. Full texts were obtained for all seven included studies, and data were extracted from the full texts. Twenty-seven records were excluded, seven studies were retained in the core corpus, and two studies were retained for the supplementary analysis (Figure 1). One study initially retained (Aprinawati & Ayu, 2026) was excluded under criterion E-6 because its results table was described by its authors as an illustrative data format whose figures could be adjusted, so the reported data could not be verified. Reasons for exclusion are detailed in Supplementary Material 2.

DOI: https://doi.org/10.17509/e.v25i3.106824
P-ISSN 0852-1190 E-ISSN 2502-0781

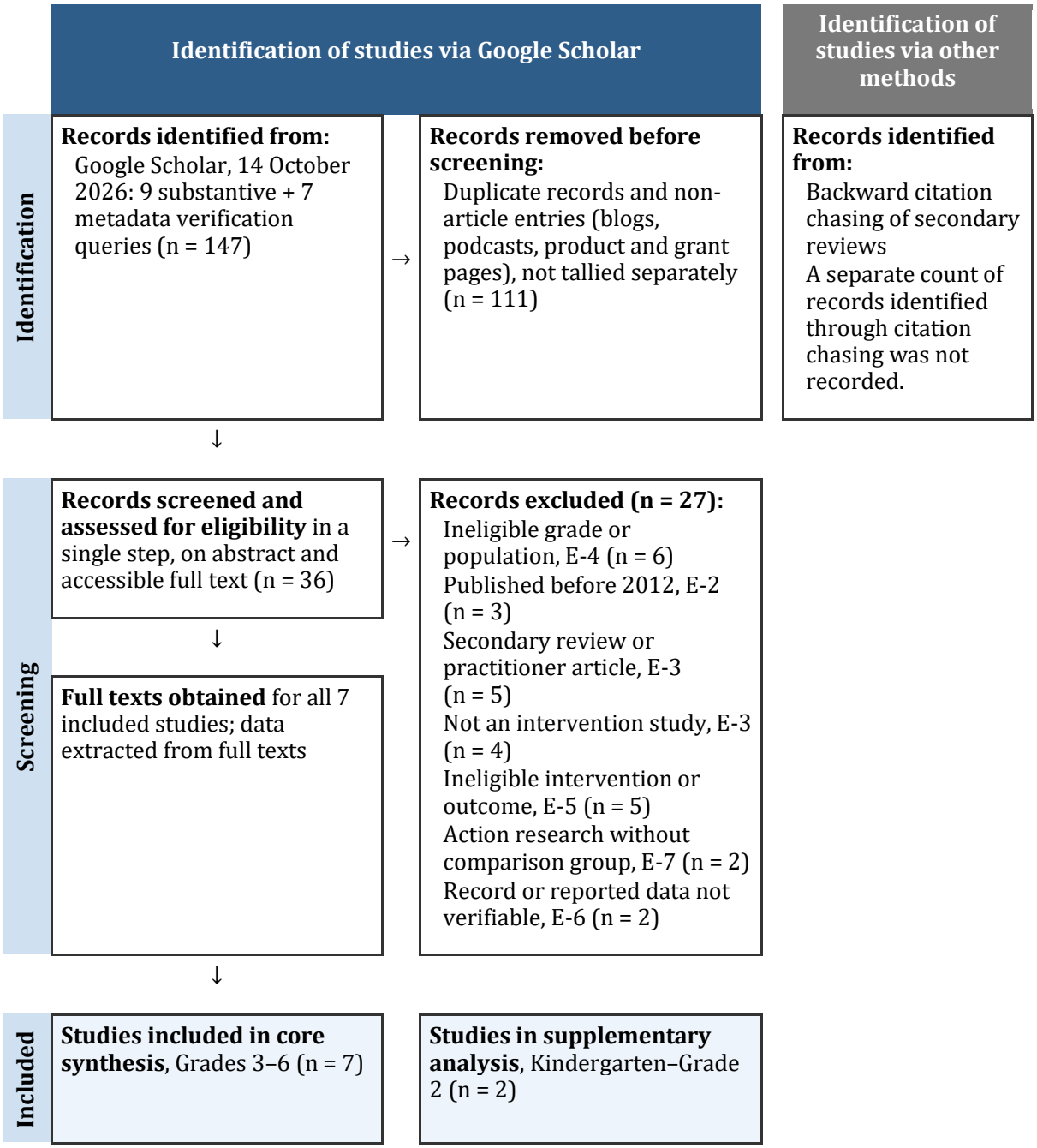


Figure 1. PRISMA 2020 flow diagram of study identification, screening, and inclusion. Title–abstract screening and full-text eligibility assessment were performed in a single step by one reviewer.

2.5. Data Extraction and Coding Framework

Data were extracted by the first author using a structured form. Deductive coding was applied along three dimensions:

1. **Six instructional components** (see Conceptual Framework), each coded as ✓ explicitly designed (the component is described and explicitly linked to the target inferential process), ◐ implicit or partial (the component is described but its link to the target process is not stated, or only part of it is described), or – not evident (the component is not reported).
2. **Five inferential thinking indicators** (see Conceptual Framework), each coded as ✓ explicitly addressed, ◐ implicit or partial, or – not evident. An indicator was coded as

DOI: https://doi.org/10.17509/e.v25i3.106824
P-ISSN 0852-1190 E-ISSN 2502-0781

addressed when it was targeted in the intervention, in the outcome measure, or in both; training and measurement were not coded separately.

3. **Text genre** (narrative, expository/informational, or mixed) and **text medium** (print, digital, or multimodal), coded as separate variables.
   Inductive coding was used to capture findings falling outside the initial categories.

2.6. Quality Appraisal and Synthesis

Methodological quality was appraised with the Mixed Methods Appraisal Tool (MMAT), version 2018 (Hong et al., 2018), using the criteria for quantitative randomized controlled trials or quantitative non-randomized studies as appropriate. Consistent with the MMAT guidance, no overall score was calculated; each study was rated high, moderate, or low on the basis of the reviewer's overall judgment across the criteria, and the intermediate ratings moderate–low and low–moderate were used when the concerns identified placed a study between two adjacent categories (the first term denoting the closer category). The principal concerns underlying each rating are reported in Table 3. Single-case designs, which the MMAT does not cover, were appraised narratively, considering the number of participants, the method of analysis, and treatment fidelity. Quality ratings were used to weight the findings and to judge the certainty of each synthesized finding rather than to exclude studies. Certainty was rated as moderate, low, or very low according to the number of contributing studies, their quality, and the consistency of their results; this is an adapted judgment rather than a formal GRADE assessment. Data were synthesized thematically (Thomas & Harden, 2008). A meta-analysis was not undertaken because of substantial heterogeneity in study designs, outcome measures, and statistical reporting. A sensitivity analysis examined whether the main conclusions changed when studies rated below moderate or the Grade 3 study were removed.

3. RESULTS AND DISCUSSION

3.1. Results

3.1.1. Study Characteristics

The seven core studies were published between 2012 and 2024 and were conducted in four countries: Indonesia (two studies), the United States (three), Norway (one), and Romania (one). Sample sizes ranged from 6 to 246 students (Table 2).

Table 2. Characteristics of the included studies.

Code / Study	Country	Design	Gr.	N (groups)	Intervention and duration	Genre / medium	Outcome measure	Main finding
S1 McMaster et al. (2012)	USA	Within-school stratified random assignment	4	246 (105/77/63)	Causal vs. general inferential vs. literal questioning in peer pairs; 3×/week, 20–30 min, 9 weeks (≈15 h)	Narrative / print	Story recall coded against causal network (κ = .89–.91)	All readers improved recall, with no condition differences overall; among struggling readers, elaborators gained more from causal (d = 0.86) and paraphrasers from general questioning (d = 1.46)

DOI: https://doi.org/10.17509/e.v25i3.106824
P-ISSN 0852-1190 E-ISSN 2502-0781

Code / Study	Country	Design	Gr.	N (groups)	Intervention and duration	Genre / medium	Outcome measure	Main finding
S2 Elbro & Buch-Iversen (2013)	Norway	School-level random assignment (16 classes, 6 schools)	6	236 (151/85)	Graphic organizers + gap-filling inference questions with teacher modeling; 8 × 30 min	Expository / print	Researcher-made inference test (α = .82); Diagnostic Reading Analysis (α = .76–.84)	Large effect on inference-making (d = 0.92) and transfer to reading comprehension (d = 0.69), sustained at 5-week follow-up
S3 Li et al. (2016)	USA	Quasi-experimental, four intact classes, counterbalanced phases	4–5	63 (15/14/16/18)	Quality Talk vs. TWA vs. hybrid; 7 mini-lessons + weekly 15-min group discussions per 6-week phase	Narrative, expository, mixed / print	Researcher-made comprehension test (κ = .81–.91); discourse analysis; oral reading fluency (AIMSweb)	Hybrid and QT effective for high-level comprehension; more authentic questions and elaborated explanations than TWA
S4 Boudreaux-Johnson et al. (2017)	USA	Single-case, alternating treatments	4	6	Close reading vs. CSR in science	Science text / NR	Daze maze task	All six students scored higher under CSR than under close reading
S5 Mih & Mih (2024)	Romania	Random assignment, small groups	3	36 (18/18)	Explanatory feedback on inference premises; 12 tutor-led sessions	Narrative / NR	Researcher-made inference test, parallel passages	Better generation and justification of inferences, not rule updating; statistics inconsistent
S6 Ardhian et al. (2020)	Indonesia	2 × 2 treatment-by-level, two intact classes	4	40 (2 classes)	DRTA vs. PQ4R; critical thinking as attribute variable	NR / print	Rubric by reading phase	DRTA > PQ4R; no technique × critical-thinking interaction
S7 Fauzia & Kelana (2020)	Indonesia	Quasi-experimental, nonequivalent groups	5	28 (2 groups)	Learning Start with a Question + online magazine	NR / digital	Literal and inferential items, total score only	Treatment mean 67.7 → 80.2; no equivalent comparison of gains

Note. NR = not reported in the primary study. Reliability is shown for S1–S3 as reported in their full texts; it was not extracted for the remaining studies. CSR = collaborative strategic reading; DRTA = directed reading–thinking activity; PQ4R = preview, question, read, reflect, recite, review; QT = Quality Talk; TWA = think before, while, and after reading.

In S1, 246 fourth graders were randomly assigned within schools to one of two inferential questioning conditions (causal or general) or a literal questioning condition; students worked in peer pairs while teachers delivered sessions of 20–30 minutes three times weekly over nine weeks (about 15 hours), with implementation fidelity of 82–90%. S2 assigned 16 sixth-grade classes in six Norwegian schools (N = 236; 151 experimental and 85 control students) to conditions by school and trained gap-filling inferences with expository texts across eight 30-minute sessions, using graphic organizers that were faded over the final sessions. S3 involved 63 fourth and fifth graders in four intact classrooms, in which teachers delivered seven mini-lessons and weekly 15-minute group discussions in each of two six-week phases. S5 involved 36 third graders with weak comprehension, divided equally between experimental and control groups, with twelve tutor-led sessions delivered under a standardized protocol. S6 compared two intact

DOI: https://doi.org/10.17509/e.v25i3.106824
P-ISSN 0852-1190 E-ISSN 2502-0781

fourth-grade classes in Cirebon Regency, whereas the comparison group in S7 completed only the posttest.

3.1.2. Quality Appraisal

Table 3. Summary of quality appraisal.

Code	Appraisal category	Quality	Principal concerns
S1	MMAT: randomized	High	Within-school stratified random assignment; groups equivalent at pretest; attrition 7%; fidelity 82–90%; subgroup analysis quasi-experimental
S2	MMAT: randomized	High	School-level random assignment of 16 classes in six schools; teachers not blind to condition; groups equivalent at pretest; at least 91% of students analyzed; proximal and standardized outcome measures; follow-up in four classes only
S3	MMAT: non-randomized	Moderate	Four intact classrooms (n = 63) with counterbalanced phases; texts not controlled for genre or difficulty; inter-rater κ = .81–.91
S4	Narrative (single-case)	Moderate–low	Single-case design, *n* = 6, visual analysis; CSR fidelity of 75.6%
S5	MMAT: randomized	Low	Within-group analysis only; degrees of freedom and effect sizes internally inconsistent
S6	MMAT: non-randomized	Low–moderate	Two intact classes; ANOVA degrees of freedom inconsistent with sample size
S7	MMAT: non-randomized	Low	Comparison group without pretest; inferential items not analyzed separately

Note. No overall MMAT score was calculated. Moderate–low and low–moderate denote studies whose concerns placed them between two adjacent categories, the first term indicating the closer category.

In S5, the within-group pretest–posttest comparison was reported as $t(34) = 2.10$ with $d = 2.97$; these degrees of freedom and this effect size are incompatible with a within-group test of 18 participants. S5's effect sizes were therefore not interpreted, and its findings were treated as low-certainty evidence.

3.1.3. RQ1: Conceptualization and Measurement of Inference

The international studies targeted specific inference types: causal inferences supporting text coherence (S1), knowledge-based inferences (S2), and text-based inferences (S5). S5 operationalized inference as the integration of two textual premises, a rule and a constraint, and required students both to answer and to justify their responses.

By contrast, the Indonesian studies tended to treat inference as a component of general comprehension. S7 divided its test items into literal and inferential questions but reported only total scores, and S6 organized its rubric by reading phase (pre-, during-,

DOI: https://doi.org/10.17509/e.v25i3.106824
P-ISSN 0852-1190 E-ISSN 2502-0781

and post-reading) rather than by inference type. Critical thinking was assessed through discourse analysis of discussions (S3) or treated as an attribute variable (S6); S6 found no interaction between reading technique and critical thinking level. S4 relied on a maze task (Daze), a general measure that does not capture the analytical processes targeted by close reading. S4 was nonetheless retained because criterion I-5 admitted general comprehension outcomes when the intervention explicitly targeted inferential or analytical processes, which close reading does. The mismatch between the processes targeted and the outcome measured is itself an alignment problem and is reflected in the coding of S4's assessment component (Table 5).

Table 4. Coverage of inferential thinking indicators.

Indicator	S1	S2	S3	S4	S5	S6	S7
Implicit meaning	◐	✓	✓	–	✓	◐	✓
Connecting information	✓	✓	◐	◐	✓	–	–
Cause and effect	✓	◐	–	–	✓	–	–
Prediction	–	–	–	–	◐	✓	–
Evidence-based conclusions	–	✓	✓	◐	✓	–	◐

Note. ✓ = explicitly addressed; ◐ = implicit or partial; – = not evident. Codes indicate whether the indicator was addressed in the intervention, the outcome measure, or both.

Prediction was the indicator least frequently addressed explicitly (one study), followed by cause and effect (two studies) and evidence-based conclusions (three studies).

3.1.4. RQ2: Strategies and Effectiveness

The certainty attached to each finding below reflects the number, quality, and consistency of the contributing studies (see Quality Appraisal and Synthesis).

Explicit inference instruction yielded the most consistent evidence (moderate certainty). In S2, a high-quality trial, the time-by-condition interaction produced a large effect on inference-making (d = 0.92) that transferred to a standardized reading comprehension measure (d = 0.69), with larger effects on interpretive than on literal questions (d = 0.73 vs. 0.45), and was sustained at a five-week follow-up in a subsample of four classes. In S5, explanatory feedback improved students' ability to generate and justify inferences but not their updating of rules and constraints; because of its internally inconsistent statistics, S5 corroborates the direction but not the magnitude of the effect, so the certainty rating rests principally on S2.

Inferential questioning had differential effects across reader subgroups (low certainty). In S1, all readers improved on story recall, and the three questioning conditions did not differ overall for struggling, average, or good readers. Effects differed, however, between two subgroups of struggling comprehenders: elaborators benefited more from causal questioning (d = 0.86), whereas paraphrasers benefited more from general questioning (d = 1.46).

Text-based discussion supported high-level comprehension (low certainty). In S3 (n = 63, four classrooms), the hybrid and Quality Talk interventions were effective, and participating students generated more authentic questions and elaborated explanations than those in the TWA condition.

DOI: https://doi.org/10.17509/e.v25i3.106824
P-ISSN 0852-1190 E-ISSN 2502-0781

Close reading showed no clear advantage (low certainty). In S4, all six students scored higher under CSR than under close reading by the end of the intervention, and annotation was the component students rated least favorably.

Indonesian studies reported positive trends but very low-certainty evidence. In S6, the DRTA group, which guided students in generating and testing predictions, achieved higher comprehension scores than the PQ4R group. In S7, the treatment group's mean rose from 67.7 to 80.2, but the design precluded an equivalent comparison of gains with the control group. No verified study examined problem-based learning for inferential outcomes in Grades 3–6.

3.1.5. RQ3: Alignment of Instructional Components

Table 5. Matrix of instructional component coverage.

Code	Objectives	Content	Model / strategy	Activities	Materials	Aligned assessment	Total ✓
S1	◐	✓	✓	✓	✓	◐	4
S2	✓	✓	✓	✓	✓	✓	6
S3	✓	◐	✓	✓	◐	✓	4
S4	✓	✓	✓	✓	◐	–	4
S5	✓	✓	✓	✓	✓	✓	6
S6	◐	–	✓	✓	–	◐	2
S7	◐	✓	✓	✓	✓	◐	4

Note. ✓ = explicitly designed; ◐ = implicit or partial; – = not evident. Total ✓ counts explicitly designed components; it describes how fully the design was specified and reported, and is not an index of intervention quality or effectiveness.

Two studies (S2 and S5) explicitly designed all six components, and both were narrowly focused, tightly controlled inference training programs. S5, for example, constructed ten passages, each containing a problem, a rule, a constraint, distracting information, and irrelevant information, and used parallel passages for the pretest and posttest. Alignment coding describes how an intervention was designed and reported, whereas quality appraisal concerns how rigorously its effects were tested; the two are therefore independent. S5 thus illustrates a well-aligned design even though the evidence it provides about effectiveness is weak. Aligned assessment was the component most often left without explicit design (explicit in three of seven studies), followed by learning objectives and instructional materials (four each). No study integrated an instructional model, a reading strategy, and explicit inference instruction within a single design: studies either trained inference in relative isolation from a broader instructional model (S1, S2, S5) or embedded comprehension within a general model or reading strategy without explicit inference instruction (S3, S4, S6, S7).

3.1.6. Supplementary Analysis: Early Grades and Digital Materials

Digital materials for inference instruction were more developed in the early grades. In a natural experiment comparing two kindergarten cohorts who completed the program before and during the COVID-19 pandemic, the ELCII system provided individualized scaffolding and feedback that supported students' inference-making (Kim et al., 2023), and van Zeijts et al. (2024) examined the stimulation of inference-making among second graders using narrative texts presented in both written and spoken form. Within the core corpus, S7 was the only study to use digital media; its students could

DOI: https://doi.org/10.17509/e.v25i3.106824
P-ISSN 0852-1190 E-ISSN 2502-0781

operate the devices independently but still required teacher guidance to access the online magazine.

3.1.7. Sensitivity Analysis

Two checks were made. First, removing the four studies rated below moderate (S4–S7) left S1–S3. The findings on explicit inference instruction, inferential questioning, and text-based discussion remained supported, by S2, S1, and S3 respectively, but the support for explicit inference instruction then rested on a single high-quality trial, which is why its certainty was rated moderate rather than high. The findings on close reading (S4) and on Indonesian studies (S6, S7) could not be evaluated without these studies and should be regarded as hypotheses. Second, removing the Grade 3 study (S5) did not change any conclusion, although it left S2 as the only study that explicitly designed all six components; restricting the corpus to Grades 4–6 therefore did not alter the overall pattern.

3.2. Discussion

The evidence base for inference instruction in the upper elementary grades rests on a small number of robust studies that explicitly train specific inference types and align their reading passages, questions, and outcome measures. This pattern is consistent with the principle of constructive alignment (Biggs, 1996): when objectives, activities, and assessment target the same cognitive process, intervention effects are more readily detected and more likely to transfer. It also accords with Elleman's (2017) meta-analysis, which documented the benefits of inference instruction for both inferential and general comprehension, and with Hall (2016), who found consistently large effects on researcher-developed inferential measures but more variable effects on standardized tests. The present review suggests one explanation for that variability: effects are more readily detected when the outcome measure is aligned with the inference type trained. S2, which trained gap-filling inferences and measured both a proximal inference task and a standardized test, shows how aligned proximal measurement can be combined with evidence of transfer. Conversely, Hall et al. (2020) reported an effect of inference instruction for Grade 6–7 English learners on a standardized comprehension test but not on their measures of inference skill, which underlines the importance of specifying which inference processes are trained and how each is assessed.

The findings also converge with Peng et al. (2023), whose network meta-analysis indicated that comprehension strategies were effective mainly when combined with background knowledge instruction: the strongest study in this review (S2) explicitly activated background knowledge for inference-making. Where this review departs from earlier syntheses is in its unit of analysis. Rather than asking whether inference instruction works, it asked how the components of each intervention were designed in relation to one another. This perspective showed that assessment, followed by objectives and materials, the components most directly under the control of teachers and curriculum developers, were the ones least often designed explicitly, a pattern that effect-size syntheses are not designed to detect.

The first implication concerns strategy selection. Although close reading is widely promoted in practice, its empirical base in elementary schools remains thin and mixed; a recent review of 19 close reading studies in Grades 3–9 likewise concluded that evidence of its effectiveness is still inconclusive, while noting that scaffolded approaches, such as those using discussion prompts, may be promising (Bogaerds-Hazenberg et al., 2026). Close reading is therefore better positioned as a reading procedure embedded within

DOI: https://doi.org/10.17509/e.v25i3.106824
P-ISSN 0852-1190 E-ISSN 2502-0781

inferential questioning, graphic representation, and explanatory feedback, the very elements that characterized the interventions with the strongest evidence.

The second implication concerns digital materials. Beyond the core corpus, a randomized trial with students in Grades 6–8 found that tutor-led instruction produced larger effects than both business-as-usual instruction (g = 0.40) and computer-delivered instruction (g = 0.30) on the inference types taught (Barnes et al., 2024). Together with the need for teacher guidance observed in S7, this finding suggests that digital materials for inferential outcomes should be designed as teacher-mediated tools rather than as substitutes for pedagogical interaction.

The third implication concerns research in Indonesia. Although the Indonesian studies reported favorable trends, they shared three recurrent weaknesses: inference was not distinguished from general comprehension, designs were weak, and statistical reporting was limited or internally inconsistent. One further Indonesian study was excluded because its reported data could not be verified.

This review identified four research gaps:

1. **Conceptual.** Indonesian studies have not distinguished between the inference types that are trained and those that are measured.
2. **Design.** No classroom design has yet integrated an instructional model, a reading strategy, and explicit inference instruction while explicitly designing all six instructional components.
3. **Technological.** Digital materials targeting inferential outcomes remain untested in the upper elementary grades.
4. **Indicator-related.** Prediction is the indicator least often addressed explicitly, followed by cause and effect.

Future design-based and development research should therefore formulate objectives around specific inference types, construct reading passages and questions with deliberately embedded inferential cues, provide scaffolding and explanatory feedback, and measure outcomes for each inferential indicator using instruments with reported reliability. Instructional models that foreground analytical thinking, such as problem-based learning, are one candidate for this integration but have not yet been tested for inferential outcomes in Grades 3–6.

3.3. Limitations

This review has several limitations that follow from its streamlined procedures and should be weighed when interpreting its conclusions.

1. The search relied on a single source, Google Scholar, supplemented by backward citation chasing; bibliographic databases such as Scopus, ERIC, Web of Science, SINTA, and Garuda were not searched. Google Scholar's ranking is not transparent and its results are not fully reproducible, duplicates and non-article entries were not tallied separately, and relevant studies, particularly those in Indonesian journals indexed in SINTA and Garuda, may have been missed.
2. Screening, eligibility assessment, data extraction, and quality appraisal were performed by a single reviewer, without inter-rater agreement testing, which increases the risk of selection and coding errors.
3. Inferential indicators were not coded separately for training and measurement, so the review cannot distinguish indicators that were taught from those that were only assessed.
4. The corpus was small (seven studies), only two studies were of high quality, and heterogeneous designs precluded meta-analysis; the certainty ratings are adapted judgments rather than a formal GRADE assessment. The review was also restricted

DOI: https://doi.org/10.17509/e.v25i3.106824
P-ISSN 0852-1190 E-ISSN 2502-0781

to peer-reviewed articles in English and Indonesian, and publication bias cannot be ruled out.

5. Alignment coding relied on what the primary studies reported, so the ratings reflect reporting transparency as well as design.
6. No formal review protocol was prepared or registered.

For these reasons, the review is best read as a structured map of the currently available evidence rather than a definitive synthesis. An update that searches bibliographic databases, uses dual independent screening, and extracts data from full texts is needed to confirm its conclusions.

4. CONCLUSION

The available evidence, though limited, suggests that inference instruction in the upper elementary grades is more likely to be effective when the inference types to be trained are specified precisely, taught explicitly through questioning, graphic representation, and explanatory feedback, and measured with aligned instruments. Few interventions explicitly design all six instructional components, and research in Indonesia requires conceptual and methodological strengthening. Integrated designs that combine an instructional model, a reading strategy, and explicit inference instruction, together with digital materials for inferential outcomes in the upper elementary grades, remain underexplored and warrant design-based and development research that attends closely to the alignment of instructional components. Given the small and uneven evidence base, these conclusions should be tested in adequately powered studies with transparent reporting.

5. AUTHORS' NOTE

The authors declare that they have no competing financial interests or personal relationships that could have influenced the work reported in this manuscript, and that the research and the writing of the manuscript adhered to the standards of research and publication ethics and are free from plagiarism. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Generative artificial intelligence and other assistive technologies were used only in a limited, supporting capacity: Claude AI assisted with brainstorming ideas, discussing concepts, outlining, and drafting or refining text, while Grammarly was used to correct grammar and spelling, edit and paraphrase the language, and improve the clarity, readability, style, and coherence of the writing. Claude AI also assisted in extracting study characteristics from the full texts of the included studies; the first author verified all extracted data against the original articles. The authors reviewed and edited all AI-assisted content to ensure its accuracy, completeness, and compliance with ethical and scholarly standards, and they take full responsibility for the final version of the manuscript. Supplementary Materials 1–3 (full query list, excluded records with reasons, and the PRISMA 2020 checklist) are submitted with this manuscript.

6. REFERENCES

Aprinawati, I., & Ayu, C. (2026). Efektivitas membaca nyaring dialogis terhadap pemahaman inferensial siswa kelas V sekolah dasar [The effectiveness of dialogic read-aloud on fifth-grade elementary students' inferential comprehension]. *Pendas: Jurnal Ilmiah Pendidikan Dasar, 11*(1), 20–42. https://doi.org/10.23969/jp.v11i01.41812

Ardhian, T., Ummah, I., Anafiah, S., & Rachmadtullah, R. (2020). Reading and critical thinking techniques on understanding reading skills for early grade students in elementary school. *International Journal of Instruction, 13*(2), 107–118. https://doi.org/10.29333/iji.2020.1328a

Barnes, M. A., Clemens, N. H., Simmons, D., Hall, C., Fogarty, M., Martinez-Lincoln, A., Vaughn, S., Simmons, L., Fall, A., & Roberts, G. (2024). A randomized controlled trial of tutor- and computer-delivered inferential comprehension interventions for middle school students with reading difficulties. *Scientific Studies of Reading, 28*(4), 411–440. https://doi.org/10.1080/10888438.2024.2331517

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education, 32*(3), 347–364. https://doi.org/10.1007/BF00138871

Bogaerds-Hazenberg, S. T. M., Droop, M., & van Steensel, R. C. M. (2026). Still too close to call: A literature review on close reading interventions and their effects on reading outcomes in Grades 3–9. *Reading Psychology, 47*(6), 898–927. https://doi.org/10.1080/02702711.2026.2644958

Boudreaux-Johnson, M., Mooney, P., & Lastrapes, R. E. (2017). An evaluation of close reading with at-risk fourth-grade students in science content. *Journal of At-Risk Issues, 20*(1), 27–35.

Cain, K., & Oakhill, J. V. (1999). Inference making ability and its relation to comprehension failure in young children. *Reading and Writing, 11*(5–6), 489–503. https://doi.org/10.1023/A:1008084120205

Chall, J. S. (1983). *Stages of reading development*. McGraw-Hill.

Elbro, C., & Buch-Iversen, I. (2013). Activation of background knowledge for inference making: Effects on reading comprehension. *Scientific Studies of Reading, 17*(6), 435–452. https://doi.org/10.1080/10888438.2013.774005

Elleman, A. M. (2017). Examining the impact of inference instruction on the literal and inferential comprehension of skilled and less skilled readers: A meta-analytic review. *Journal of Educational Psychology, 109*(6), 761–781. https://doi.org/10.1037/edu0000180

Fauzia, N. L. U., & Kelana, J. B. (2020). Meningkatkan kemampuan membaca pemahaman siswa dengan media majalah online menggunakan model kooperatif Learning Start with a Question di kelas V SD [Improving fifth-grade students' reading comprehension through online magazines and the cooperative Learning Start with a Question model]. *COLLASE (Creative of Learning Students Elementary Education), 3*(4), 174–181.

Garritty, C., Gartlehner, G., Nussbaumer-Streit, B., King, V. J., Hamel, C., Kamel, C., Affengruber, L., & Stevens, A. (2021). Cochrane Rapid Reviews Methods Group offers evidence-informed guidance to conduct rapid reviews. *Journal of Clinical Epidemiology, 130*, 13–22. https://doi.org/10.1016/j.jclinepi.2020.10.007

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*(3), 371–395. https://doi.org/10.1037/0033-295X.101.3.371

Hall, C. S. (2016). Inference instruction for struggling readers: A synthesis of intervention research. *Educational Psychology Review, 28*(1), 1–22. https://doi.org/10.1007/s10648-014-9295-x

Hall, C., Vaughn, S., Barnes, M. A., Stewart, A. A., Austin, C. R., & Roberts, G. (2020). The effects of inference instruction on the reading comprehension of English learners with reading comprehension difficulties. *Remedial and Special Education, 41*(5), 259–270. https://doi.org/10.1177/0741932518824983

Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., Vedel, I., & Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information, 34*(4), 285–291. https://doi.org/10.3233/EFI-180221

Kementerian Pendidikan dan Kebudayaan. (2016). *Peraturan Menteri Pendidikan dan Kebudayaan Nomor 22 Tahun 2016 tentang Standar Proses Pendidikan Dasar dan Menengah* [Regulation of the Minister of Education and Culture No. 22 of 2016 on process standards for primary and secondary education].

Kim, J., Burey, J., Hwang, H. J., McMaster, K., & Kendeou, P. (2023). Supporting inference-making during COVID-19 through individualized scaffolding and feedback: A natural experiment. *Reading and Writing, 36*(2), 467–490. https://doi.org/10.1007/s11145-022-10391-2

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Kispal, A. (2008). *Effective teaching of inference skills for reading: Literature review* (Research Report DCSF-RR031). National Foundation for Educational Research.

Li, M., Murphy, P. K., Wang, J., Mason, L. H., Firetto, C. M., Wei, L., & Chung, K. S. (2016). Promoting reading comprehension and critical–analytic thinking: A comparison of three approaches with fourth and fifth graders. *Contemporary Educational Psychology, 46*, 101–115. https://doi.org/10.1016/j.cedpsych.2016.05.002

McMaster, K. L., van den Broek, P., Espin, C. A., White, M. J., Rapp, D. N., Kendeou, P., Bohn-Gettler, C. M., & Carlson, S. (2012). Making the right connections: Differential effects of reading intervention for subgroups of comprehenders. *Learning and Individual Differences, 22*(1), 100–111. https://doi.org/10.1016/j.lindif.2011.11.017

Mih, V., & Mih, C. (2024). Text-based inference instruction for elementary grade children with reading comprehension difficulties: An intervention research. *Studia Universitatis Babeș-Bolyai Psychologia-Paedagogia, 69*(1), 257–272. https://doi.org/10.24193/subbpsyped.2024.1.13

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects*. Authors.

OECD. (2023a). *PISA 2022 assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/dfe0bf9c-en

OECD. (2023b). *PISA 2022 results (Volume I): The state of learning and equity in education*. OECD Publishing. https://doi.org/10.1787/53f23881-en

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Peng, P., Wang, W., Filderman, M. J., Zhang, W., & Lin, L. (2023). The active ingredient in reading comprehension strategy intervention for struggling readers: A Bayesian network meta-analysis. *Review of Educational Research*. Advance online publication. https://doi.org/10.3102/00346543231171345

Taba, H. (1962). *Curriculum development: Theory and practice*. Harcourt, Brace & World.

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology, 8*, 45. https://doi.org/10.1186/1471-2288-8-45

Tyler, R. W. (1949). *Basic principles of curriculum and instruction*. University of Chicago Press.

van Zeijts, B. E. J., Ganushchak, L. Y., de Koning, B. B., & Tabbers, H. K. (2024). Stimulating inference-making in second grade children when reading and listening to narrative texts. *Reading and Writing, 37*(9), 2201–2227. https://doi.org/10.1007/s11145-023-10463-x

Wanzek, J., Wexler, J., Vaughn, S., & Ciullo, S. (2010). Reading interventions for struggling readers in the upper elementary grades: A synthesis of 20 years of research. *Reading and Writing, 23*(8), 889–912. https://doi.org/10.1007/s11145-009-9179-5

Wiggins, G., & McTighe, J. (2005). *Understanding by design* (2nd ed.). ASCD.

DOI: https://doi.org/10.17509/e.v25i3.106824
P-ISSN 0852-1190 E-ISSN 2502-0781