History Education Research Journal

History Education Research Journal

UCLPRESS

Systematic Reviews

# A systematic review of pedagogical interventions on the learning of historical literacy in schools

Kim Wilson,[1,*] Dean Dudley,[2] Janet Dutton,[1] Renee Preval-Mann,[3] Elizabeth Paulsen[4]

[1] Senior Lecturer, Macquarie University School of Education, Sydney, Australia
[2] Associate Professor, Macquarie University School of Education, Sydney, Australia
[3] PhD candidate and Sessional Academic, Macquarie University School of Education, Sydney, Australia
[4] Sessional Academic, Macquarie University School of Education, Sydney, Australia
* Correspondence: kim.wilson@mq.edu.au

Submission date: 19 June 2023; Acceptance date: 26 October 2023; Publication date: 6 December 2023

**How to cite**
Wilson, K., Dudley, D., Dutton, J., Preval-Mann, R. and Paulsen, E. (2023) 'A systematic review of pedagogical interventions on the learning of historical literacy in schools'. *History Education Research Journal*, 20 (1), 9. DOI: https://doi.org/10.14324/HERJ.20.1.09.

**Peer review**
This article has been peer-reviewed through the journal's standard double-anonymous peer-review process, where both the reviewers and authors are anonymised during review.

## Abstract

Over the past thirty years, there has been a growing body of research investigating the efficacy of pedagogical interventions to enhance the historical literacy skills of primary and secondary school students. However, there exists no systematic review or meta-analysis summarising the impact of such research or the efficacy of interventions trialled. The purpose of this systematic review is to identify pedagogies that have a demonstrable effect on students' historical literacy skills, with a particular interest in those pedagogies that have a measurable positive effect on historical epistemological knowledge and skills. Findings of this review indicate that when a discrete historical epistemological knowledge or skill is targeted by a pedagogical intervention that utilises a discipline-specific scaffolded heuristic, there is greater likelihood of positive outcomes for student learning. However, the significant heterogeneity between studies, and the diversity in the comparisons being made by the included studies, make it difficult to identify the most effective intervention. This systematic review establishes the characteristic features of pedagogical historical literacy interventions from the available research reporting credible findings.

**Keywords** history; teaching; pedagogy; historical literacy; historical thinking

# Introduction

Over the past thirty years, there has been a growing body of research investigating the efficacy of pedagogical interventions to enhance the historical literacy skills of primary and secondary school students. Non-systematic exploration of effective historical classroom pedagogies, for example, Nokes and De La Paz (2023: 350) investigated historians' reading heuristics and procedures, and noted the difficulty of fostering students' historical argumentation, especially when 'processes are contingent upon fragile and emerging understandings of the nature of history as a discipline'. Luís and Rapanta (2020) conducted a thorough review into how historical reasoning competence has been operationalised in history education empirical research. They found a clear predominance of studies focusing on content knowledge acquisition skills, together with a lack of empirical research investigating the full suite of historical reasoning competence skills (Luís and Rapanta, 2020: 10–11). However, there exists no systematic review or meta-analysis summarising the impact of such empirical research or the efficacy of interventions trialled. The purpose of this systematic review is to identify pedagogies that have a demonstrable effect on students' historical literacy skills, with a particular interest in those pedagogies that have a measurable positive effect on historical epistemological knowledge and skills. We are also interested in collating information about pedagogies that can be transferred from experimental conditions to a mixed-ability primary or secondary classroom. The research question that guided the review was: *What is the relationship between pedagogical strategies and improved historical literacy in primary and secondary school children?*

## Historical literacy: what does it mean?

There is great diversity in the terminology used to describe historical literacy. Terminology ranges from basic historical recall and narration, to more sophisticated explanation, analysis and evaluation. We provide characteristic features of the language used to describe historical literacy, grouped into two strands: (1) historical content knowledge; and (2) historical epistemological knowledge and skills.

(1) *Historical content knowledge*

Historical content knowledge may be referred to as factual, historical or objective knowledge. The acquisition of historical content knowledge is typically demonstrated through description, narration, factual recount or recall in response to comprehension-style questions. This type of knowledge acquisition may be referred to as concrete or lower-order thinking, because the internalisation of historical knowledge maps to what Krathwohl's (2002) Taxonomy of Educational Objectives refers to as Remembering (recognising and recalling) and Understanding (determining the meaning of instructional messages, especially classifying and summarising).[1] Researchers refer to the attainment of historical content knowledge through processes of: memorisation (Aidinopoulou and Sampson, 2017); use of historical vocabulary, sequencing events and periods, and identifying characteristic features (de Groot-Reuvekamp et al., 2018); or, for example, as the successful acquisition of background knowledge on a given historical topic (Wissinger et al., 2021).

(2) *Historical epistemological knowledge and skills*

Historical epistemological knowledge and skills can be divided into two main categories: *deconstruction* of source material, and *reconstruction* of historical narrative or argument. We drew on a selection of studies included in this systematic review (Ariës et al., 2015; Bertram et al., 2017; De La Paz and Felton, 2010) to define the elements of source *deconstruction* and historical narrative/argument *reconstruction*:

- *Deconstruction* – critical analysis of historical sources to ascertain context, audience, message; purpose of source creation and perspective represented; techniques used to communicate the message, purpose and perspective of a historical source.
- *Reconstruction* – interpretation, reasoning and explanation of historical evidence; analysis and synthesis of evidence and historical argument; analysis and reasoning leading to a judgement expressed as an assessment of value or an evaluation based on criteria.

The development of historical epistemological knowledge and skills is often referred to as abstract or higher-order thinking, because the demonstration of these skills and knowledge maps to what Krathwohl's (2002) Taxonomy of Educational Objectives refers to as Applying (executing or implementing), Analysing (identifying constituent parts and detecting how parts relate to one another), Evaluating (making judgements) or Creating (combining elements to make an original product). Researchers refer to the development of historical epistemological knowledge and skills as, for example: sourcing, substantiation, corroboration, perspective, contextualisation and rebuttal (De La Paz and Felton, 2010; Huijgen et al., 2018; Nokes et al., 2007; Reisman, 2012; Wissinger and De La Paz, 2016). Most pedagogical empirical studies in secondary and primary school history contexts focus on an intervention to improve some aspect of students' historical literacy in epistemological knowledge and skills, possibly because the demonstration of abstract thinking is a key feature of academic achievement at the highest levels.

Our systematic review seeks to demonstrate the efficacy of historical literacy pedagogical interventions in improving student historical content knowledge and/or historical epistemological knowledge and skills. Following, we provide a detailed overview of all steps and processes taken in this systematic review, including detailed notations on reasons for study inclusion and exclusion, and the methods of critically appraising included studies. Our review is informed by a clear theory of change regarding historical literacy education, and the synthesis of data appraises pedagogical interventions in terms of feasibility, replicability, extent of academic gain, and explicit or implicit moderating variables affecting intervention results. Implications of findings are discussed, and recommendations for further research are suggested.

## Methods

A systematic literature search was conducted using electronic databases (PsychINFO, ERIC, Academic Search Premier, Education Research Complete, Humanities International Complete) from 1990 to February 2023. The search protocol guided the database search (see Appendix A). Included studies reported on at least one of two primary outcomes (PO): PO(i) historical recount or historical description or historical narrative; and PO(ii) historical explanation or historical interpretation or historical judgement – in combination with accurate historical knowledge. The search was limited to peer-reviewed journal articles written in English.

Search terms included the use of three broad categories: (1) historical content knowledge; (2) historical epistemological knowledge and skills; and (3) educational context. Searches used the following terms:

1. (histor* n2 (teach* OR literac* OR knowlege* OR curriculum OR instruct* OR descript* OR interpret* OR narrative* OR source analys*)) and
2. ((primary OR elementary OR junior OR senior OR secondary OR high OR grammar OR grade) n2 school* OR freshman OR sophomore OR grade* OR high school students OR secondary education OR schools OR high school teachers)

### Study selection, data collection process and data items

Included studies took place with school-aged children (about 5–18 years old), were delivered in a school classroom to the whole class group (either by the classroom teacher or by a guest instructor) and in a regular school. We selected studies conducted under these conditions because we were interested in identifying historical literacy pedagogical strategies that could be scaled up for large cohorts, and potentially delivered state-wide. Studies were selected on pedagogical interventions that were conducted over a sustained period (one week or more, with a minimum of three sequential lessons), and were empirical, reported on at least one primary outcome and investigated the use of a historical literacy pedagogical strategy.

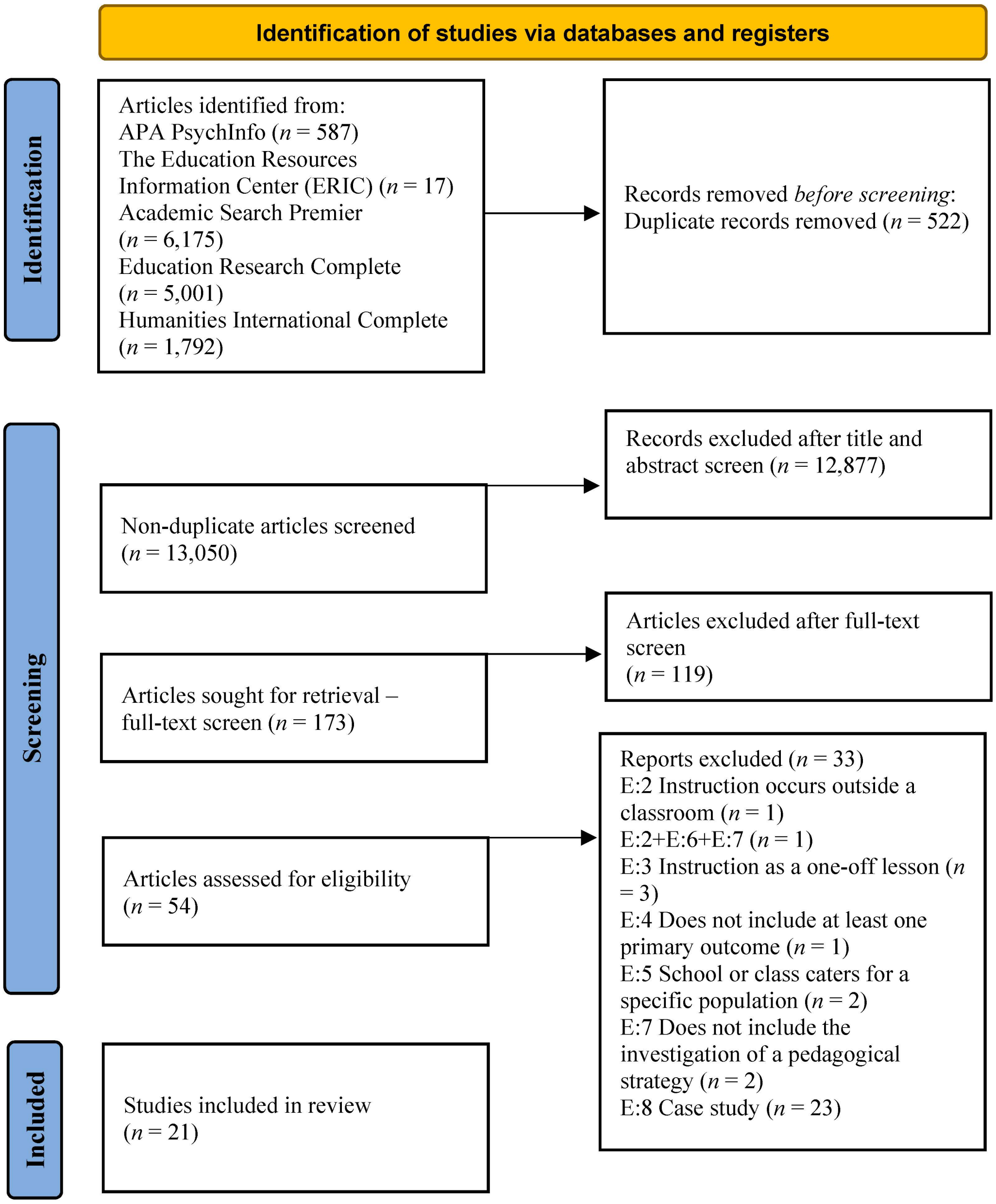
Studies were excluded if they were delivered at university or outside a regular school classroom (for example, in a museum or to a withdrawn group of students), or if instruction was provided as a one-off lesson or presentation. Furthermore, studies were excluded if they were not empirical, they did not report on at least one primary outcome, or there was no investigation of a historical literacy pedagogical intervention.

Three researchers (R1, R2 and R3) worked on the selection of studies. R1 searched databases for relevant literature. Following the removal of duplicates, R1 and R2 completed a full title sweep of the 13,050 articles independently. If the researcher was unsure whether to include an article based on title, the abstract was consulted; if still unsure, the article was included. In total, 173 articles were agreed upon as meeting the inclusion criteria at this point, and proceeded for full text review. R2 and R3 completed reads of full articles independently, and mutually agreed on the exclusion of 119 articles. The remaining 54 articles were identified for possible inclusion and assessed for eligibility. R2 completed detailed reads of each, and recorded reasons for any study being excluded in this phase. This resulted in a further 33 studies being excluded, and allowed 21 studies to be included in this review. See Figure 1 for an overview of the study selection process.

R3 extracted data from the 21 included studies, and R2 checked a 20 per cent sample for accuracy; no anomalies were found. The data extraction included: inclusion criteria, evidence hierarchy; aim/objective or focus of study; research question(s); study design; recruitment, participants, study setting, pedagogical strategy/intervention; findings.

Once this data extraction had occurred, the fourth researcher (R4) was consulted on whether a meta-analysis of the included studies was possible and/or appropriate. R4 entered the statistical data from all 21 studies into the Comprehensive Meta Analysis software (version 4.1). Heterogeneity was assessed across all the studies using a series of complementary statistical analyses, such as the Q statistic, inconsistency index ($I^2$), Tau statistic ($T^2$) and prediction interval. R4 determined that based on the significant heterogeneity present in the included studies, and the diversity in the comparisons being made by the primary studies, it would be inappropriate to combine all included studies in a single meta-analysis, given the mix of comparisons of different pedagogical interventions with different outcome variables. Furthermore, a meta-analysis of studies that present risk of bias is likely to generate an effect size that is misleading.

**Figure 1. PRISMA flowchart of identified studies**


Identification of studies via databases and registers
Identification
Articles identified from:
APA PsychInfo (n = 587)
The Education Resources Information Center (ERIC) (n = 17)
Academic Search Premier (n = 6,175)
Education Research Complete (n = 5,001)
Humanities International Complete (n = 1,792)
Records removed before screening:
Duplicate records removed (n = 522)
Screening
Non-duplicate articles screened (n = 13,050)
Records excluded after title and abstract screen (n = 12,877)
Articles sought for retrieval – full-text screen (n = 173)
Articles excluded after full-text screen (n = 119)
Articles assessed for eligibility (n = 54)
Reports excluded (n = 33)
E:2 Instruction occurs outside a classroom (n = 1)
E:2+E:6+E:7 (n = 1)
E:3 Instruction as a one-off lesson (n = 3)
E:4 Does not include at least one primary outcome (n = 1)
E:5 School or class caters for a specific population (n = 2)
E:7 Does not include the investigation of a pedagogical strategy (n = 2)
E:8 Case study (n = 23)
Included
Studies included in review (n = 21)

## Critical appraisal

Two critical appraisal models suitable for evaluating qualitative studies were consulted. Our Critical Appraisal and Weight of Evidence (WoE) template (see Appendix B) was designed with reference to the Critical Appraisal Skills Programme (CASP) template and the Cochrane Effective Practice and Organisation of Care (EPOC) protocol and template. These templates have been used in similar qualitative systematic reviews (see Pino and Mortari, 2014; Sterman et al., 2016).

The Critical Appraisal and WoE template was completed independently by R2 and R3 for every included study. Each study was assessed for internal methodological coherence, with criteria including: clarity of research question or aim; study and sample design; setting, participants and recruitment; data collection and analysis procedures; traceability of research processes; and inclusion of researcher background or orientations. Implicit reference to sample design was accepted. Criteria were scored on a *yes* (1 point) or *no/cannot tell* (0 point) measure. The numerical score for internal methodological coherence was mapped to a 5-point value scale: *high* (9–10), *high-medium* (7–8), *medium* (5–6), *medium-low* (3–4) and *low* (1–2). Each study was further assessed for its relevance to the review question, with criteria including: description of pedagogical intervention (*detailed* – 2 points; *general* – 1 point; *no description* – 0 points); a defined historical literacy skill (*yes* – 1 point; *no* – 0 points); and primary outcomes reported (*PO(ii)* – 2 points; *PO(i)* – 1 point). A detailed description of the pedagogical intervention was one that could be replicated by an expert teacher practitioner (R2, an experienced secondary school history teacher and university history education lecturer, made the final judgement on detailed versus general description of the pedagogical intervention). The numerical score for the relevance of the studies to the review question was mapped to a 3-point value scale: *high* (5), *medium* (3–4), *low* (1–2). Results for the internal methodological coherence and relevance to review question appraisal were combined to report a WoE 5-point value scale: *high* (13–15), *high-medium* (10–12), *medium* (7–9), *medium-low* (4–6), and *low* (1–3). R2 compared the completed templates for each study, and resolved any differences via direct reference to the study. Where relevant, page references were noted in support of the resolution. Results of the Critical Appraisal and WoE are provided in Table 1.

### Synthesis methods

Informing our review was a clear theory of change regarding historical literacy education. We hypothesised that when a discrete historical literacy pedagogical intervention was implemented in a primary or secondary school classroom, students would learn a discrete history thinking skill. The acquisition of the discrete history thinking skill would: (1) be apparent in students' reuse of the learnt skill in other similar circumstances; and (2) lead to improved historical thinking, evident in an academic gain demonstrated through a specified measurement tool (for example, an essay or test) with no detrimental effect on students' acquisition of required historical content knowledge. Our synthesis of data follows Weiss's (1997) theory-based evaluation method, whereby we set out to appraise the:

- nature of historical literacy pedagogical interventions in terms of being fit for purpose
- steps taken in implementations of pedagogical interventions
- feasibility and replicability of historical literacy pedagogical interventions
- nature and extent of academic gains linked to interventions
- explicit or implicit moderating variables that may affect intervention results.

Hence, our synthesis seeks to demonstrate how historical literacy pedagogical interventions work, why they work and for whom the interventions work.

## Results of systematic review

Studies are referred to in this section by number assigned according to alphabetical ranking. See Table 1 for number ranking (#) and author(s). Of the 21 studies reviewed, 4 were conducted in primary school contexts (#1, 5, 10 and 21), and 17 were conducted in high school contexts (#2–4, 6–9 and 11–20). Interventions were delivered over a period of one week to one school year, and focused on historical content relevant to each study context. Smaller scale studies were comprised of participants from one or two experimental class(es) with corresponding control class(es) (for example, #1, 2, 5, 12 and 18). Large-scale studies involved multiple year groups (#13, 21), multiple school sites (#4, 10, 17 and 19), and/or substantial participant recruitment (#4, $n$ = 900; #7, $n$ = 1,330; #8, $n$ = 1,029; #10, $n$ = 788; #19, $n$ = 1,022; #21, $n$ = 608).

**Table 1. Description of studies and key findings (*n* = 19) *WoE maximum score = 15**

| # | Authors, year/country | Participants | Method of data collection | Method of data analysis | Quality control | Evidence hierarchy (EH) and WoE | Results |
|---|---|---|---|---|---|---|---|
| 1 | (Aidinopoulou and Sampson, 2017)<br>Greece | **49** Grade 5 students | Teacher logs.<br>Post-test. | The Assessing Historical Thinking and Understanding (ARCH): Historical Thinking Skills (HTS) Rubric was used.<br>The Mann–Whitney U test was used to calculate the effect size, to examine the significance of differences in students' HTS achievement scores between experimental and control groups. | Post-test: Historical content assessed with the standardised tests of the National Curriculum.<br>Mann–Whitney U tests were employed to investigate for potential significant differences in the assessment scores between the two groups, since the data did not follow a normal distribution and the sample size was not large. | EH-5<br>Credibility and dependability of findings are vulnerable due to a lack of identified quality controls. Study is relevant to review question, but description of intervention is generalised.<br>**WoE: *medium-high* (10)*** | Flipped classroom had no discernible effect on students' acquisition of historical content knowledge; however, gains are reported for the experimental group in historical epistemological knowledge and skills. |
| 2 | (Ariës et al., 2015)<br>Netherlands | Experiment 1: **92** Grade 10 students<br>**3** teachers<br>Experiment 2: **63** Grade 10 students<br>**3** teachers | Experiment 1: Two pre-tests, and one intermediate, and post-test.<br>Experiment 2: Pre-test, intermediate test and a post-test. | Experiment 1: Paired *t*-tests to measure post-test scores. Independent *t*-tests to compare experimental and control groups – Kolmogorov–Smirnov test to identify any deviation from normality.<br>Experiment 2: Kolmogorov–Smirnov test used to measure for group variances in pre-test and post-test scores for reasoning achievement. | Experiment 1: Results were compared with control group data to control for a natural increase of reasoning skills.<br>Experiment 2: Pre-test analysis accounted for homogeneity of both groups. | EH-5<br>Dependability of study design, data collection and analysis acceptable, but lack of detail in historical content and question types threatens credibility of conclusions. Study is relevant for the review question, but intervention is not replicable.<br>**WoE: *medium-high* (12)** | Experimental group reported with improved test results as a consequence of better working memory capacity and internalisation of reasoning structures. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 3 | (Azor et al., 2020)<br>Nigeria | **70** Senior secondary students<br>**4** teachers | Pre- and post-tests. | History Achievement Test (HAT) scored out of 30, with correct answers awarded 1 mark and incorrect answers awarded 0. History Interest Inventory Scale (HIIS) scored on 4-point scale of agreement. Data analysed using SPSS. Mean and standard deviation used to answer research questions. ANCOVA used to test null hypotheses at 0.05 significant level. | Instruments were validated for content credibility by 3 experts in different departments from same university as lead researcher. Trial testing of instruments carried out for reliability. Kuder–Richardson Formula 20 (KR20) used to test achievement scores, which yielded reliability index 0.65. Cronbach's alpha used to test reliability of HIIS. | EH-5<br>Credibility and dependability of study design, data collection, data analysis and quality control is strong. Relevance for the review question is *low* – no description of intervention and HL skill not clearly identified.<br>**WoE:** ***medium-high*** **(10)** | Authors note a statistically significant difference between the mean achievement score of students taught with the intervention and those taught without. |
| 4 | (Bertram et al., 2017)<br>Germany | **900** Grade 9 students | Pre-test, post-test and follow-up maintenance test. | Missing data were multiply imputed with IVEWare. Data from the total sample were used for scaling. Competence tests and scores were estimated in separate uni-dimensional one-parameter (partial credit) item response theory models with the software ConQuest. All variables measure at each of the 3 measurement points. All analyses were computed in SAS with PROC SURVEYREG and PROC MIANALYZE. | The same teacher taught the 30 intervention classes. Every lesson was observed by two raters who assessed the interactions between the teacher and the class. An observation protocol was designed based on teaching quality studies, and included 6 scales assessing for, e.g., discipline, use of time, students' cooperation and teacher's clarity. Randomisation checks and treatment checks performed. | EH-4<br>Credibility and dependability of study design, data collection, data analysis and quality control were all *high*. Relevance for the review question was *high*.<br>**WoE:** ***high*** **(15)** | Intervention groups scored better than control groups on 4 of 5 achievement tests, and learned the same or more in terms of historical content.<br>However, students in the intervention group with the live historical source scored lower on two tests (understanding deconstruction, and understanding oral history). |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 5 | (Brugar, 2016)<br>USA | **50** Grade 5 students<br>**3** teachers | Teacher interviews and classroom observations (4x per/wk).<br>Pre- and post-assessments. | Paired sample *t*-tests to compare pre- and post-assessments for experimental group. Sample *t*-tests to compare students' gains scores. Cohen's *d* for independent sample *t*-tests for effect size.<br>Three-step interpretative approach for thematic analysis of qualitative data. | Two forms of the assessment (A&B) developed to lessen threat to internal validity or potential practice effects.<br>Assessments and rubrics reviewed by an experienced teacher. Assessments field-tested with 8 non-participating students.<br>A&B Assessments scored independently, and then compared – inter-rater reliability was 96% (A) and 94% (B), respectively.<br>No quality control for qualitative thematic analysis of data. | EH-5<br>Data analysis, quality control, credibility and dependability of findings are vulnerable due to a lack of identified quality controls for Qual data.<br>Study is relevant to review question, but description of intervention is generalised.<br>**WoE: *medium-high* (11)** | Modest gains in assessment scores for experimental group.<br>Author reports experimental condition used disciplinary strategies of observation, sourcing and contextualising, but no comparative comment with comparison group. |
| 6 | (De La Paz and Felton, 2010)<br>USA | **160** Grade 11 students<br>**4** teachers | Pre- and post-tests. | Essays scored in a multi-stage process for the development of arguments using Toulmin's (1958) model for interpreting arguments.<br>Stage 1: essays coded to identify all claims for and against a position. Stage 2: each claim coded for level of development on a 4-level scale. Stage 3: claims that opposed the writer of the source were coded for the degree to which they were responded to in rebuttal.<br>Pre-test scores used as covariate for effectiveness of intervention at post-test.<br>Ordinal regression for all ranked measures. Repeated-measures ANOVA to explore results for 2 measures at post-test. | Experimental teachers and an American history professor reviewed all material. Revised materials were used for data collection.<br>Treatment validity included regular lesson observations and weekly communications. Field notes were compared with written lesson plans.<br>Two graduate students completing their single subject credentials in social sciences scored the essays for overall persuasiveness and historical accuracy. IRR was .90.<br>Author developed coding schemes and a 25% random sample scored independently by a trained assistant. | EH-5<br>Credibility and dependability of study design, data collection, data analysis and quality control is *high*.<br>Relevance for the review question is *high*.<br>**WoE: *high* (14)** | Integration of disciplinary reading and writing strategies for poor and average high school writers' argumentative essays had a positive impact on their ability to construct and support a historical argument.<br>Students in the experimental group wrote argumentative essays with more advanced development of claims and rebuttals after instruction.<br>Students in the experimental group documented citations and quotations to further their arguments more frequently than control group in the post-test. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 7 | (De La Paz et al., 2014)<br>USA | **1,330** Grade 8 students<br>**13** teachers | Pre-test and post-tests.<br>Observer field notes, fidelity of implementation protocols and teacher reflections. | A representative subset of **310** students from each condition selected. Students' historical essays were analysed using 3 measures: historical arguments (analytic rubric), holistic quality (holistic rubric) and essay length (number of words, regardless of spelling).<br>Data were analysed using hierarchical linear models because the students were nested within classes. | Fidelity checks were conducted.<br>Historical argument: 2 pairs of raters used the analytic rubric to score pre- and post-test essays, IRR for substantiation, .87; IRR for perspective, .93; IRR for contextualisation, .88.<br>Holistic quality: clarity and persuasiveness of students' response rated 0–6. Raters scored complete set of essays with 93% agreement.<br>Essay length: scored by independent raters with a sample of 50 papers counted twice. | EH-5<br>Credibility and dependability of study design, data collection, data analysis and quality control is *high*.<br>Study relevance for the review question is *medium*, with a generalised description of the intervention.<br>**WoE**: ***high*** **(14)** | Results report that scaffolded instruction on high-school students' discipline-specific reading and writing has a positive impact on their ability to construct historical arguments.<br>Cognitive apprenticeship model found effective for readers at the highest proficiency levels, together with those who struggle academically. |
| 8 | (De La Paz et al., 2017)<br>USA | **1,029** Grade 8 students<br>**36** teachers | Pre- and post-test argumentative Writing Tasks, Teacher and Student Fidelity to the intervention checks. | Hierarchical linear modelling. Two-level random intercept models, with students at Level 1 and teachers at Level 2, to examine the effects of participating in the disciplinary writing curriculum intervention on three aspects of students' disciplinary writing skills: historical reasoning, writing quality and essay length. | Estimated the models using restricted maximum likelihood estimation. All student-level variables were grand mean centred in all analyses. The Level 1 intercept, therefore, is the average disciplinary writing skills of students net of differences among teachers in their students' characteristics. | EH-5<br>Credibility and dependability of study design, data collection, data analysis and quality control were all *high*.<br>Relevance for the review question was *high*.<br>**WoE**: ***high*** **(15)** | Results report the curriculum intervention and teacher PD resulted in improved historical writing and general argument writing for diverse learners. Results also indicate effectiveness of the cognitive apprenticeship approach for readers at higher proficiency levels and those who struggled academically. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 9 | (De La Paz, 2005)<br>USA | **132** Grade 8 students | Pre- and post-tests for experiment group.<br>Post-test for control group. | Essay length: essays scored for number of words written.<br>Persuasive quality: holistic persuasive rating scale used by 2 teachers who rated each essay from *low* (0) to *high* (6).<br>Number of arguments: each essay argument segmented and categorised into following elements – claim, rebuttal, alternative solution, countered rebuttal, justification, warrant and constraint.<br>Historical accuracy: holistic rating scale to assess extent to which the writer used facts accurately, whether facts related to the historical question, and extent to which facts were used to create context for writer's argument.<br>Historical understanding: 25 experimental students interviewed about their understanding of ways to think about history and historical thinking. 3-point scale scoring rubric used to code responses. | Experimental teachers reviewed 2 or more drafts of all materials. Revised and approved materials were used for data collection. Treatment validity checks were conducted.<br>Essay length: an undergraduate and graduate student counted a topic set each and a 25% random sample of the alternate set.<br>Persuasive quality: Raters provided with benchmarked essays, and told to ignore factors not related to persuasiveness. Student essays were handwritten, and any essay with poor spelling was recopied by an assistant.<br>Number of arguments: ¼ of essays independently scored by a teacher unfamiliar with the project.<br>Historical accuracy: Raters were provided with benchmarked essays, and told to ignore factors not related to historical accuracy.<br>Historical understanding: Interviews independently scored by author and assistant | EH-5<br>Credibility and dependability of study design, data collection, data analysis and quality control is *medium-high*.<br>Relevance for the review question: *high*.<br>**WoE: *high* (13)** | Experimental group wrote significantly better essays than students who did not receive the intervention.<br>Effect sizes were greatest for essay length, persuasiveness, and number of arguments.<br>Effect size for accuracy of historical content low, but statistically significant.<br>The 25 students interviewed about their historical understanding and historical thinking showed modest yet favourable findings regarding their development of historical reasoning.<br>Results reported that after instruction, students wrote longer and more persuasive essays containing more arguments and more accurate historical content. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 10 | (de Groot-Reuvekamp et al., 2018)<br>Netherlands | **788** Grade 2 and Grade 5 students<br>**16** teachers | Pre- and post-tests. 14 questions common to Grades 2 and 5. Same test used for pre- and post-test. | Separate analyses for Grades 2 and 5, standardised score on historical time as the dependent variable. Step 1: three-level null model (Model 0) was estimated without explanatory variables, and used to determine the variance within and between classes, before considering differences between conditions. Step 2 of analysis included explanatory variables. Model 1 – standard 3-level mixed effects model with fixed effects for time, condition, and their interaction. | Pre- and post-test validated in a prior study (de Groot-Reuvekamp et al., 2018). Treatment fidelity was determined through observations and questionnaires – 16 observations rated by first author, with remainder rated by a teacher trainer from a different faculty. Fidelity checks conducted (and demonstrated 6 teachers in experimental condition spent an additional 24 mins per/wk on history lessons). | EH-5<br>Dependability of study design and data collection is strong. However, fidelity check issue threatens data analysis and quality control. Study is relevant for the review question, but reports PO(i) only, and description of intervention is generalised.<br>**WoE**: ***medium-high*** **(12)** | Grade 2: greater progression in students' understanding of historical time in experimental groups. However, gains also made in control group with no history lessons. Grade 5: students in experimental group scored significantly higher on the post-test. However, 6 teachers in the experimental condition spent an additional 24 minutes weekly on history, providing their students with both the intervention lesson and the traditional lesson. |
| 11 | (Del Favero et al., 2007)<br>Italy | **100** Grade 8 students | Two pre- and post-tests. Two open-ended questionnaires. Situational Interest (SI) questionnaire. | Principal component analysis (PCA) with varimax rotation and Rasch models used to assess dimensionality of Situational Interest questionnaire. Logit scores for each factor used for the SI questionnaire. Cohen's kappa used for reliability indexes. | 30% of pre- and post-tests scored independently by 2 judges. Open-ended questionnaire testing students' historical problem solving scored independently by two judges. Two independent judges attributed answers for the topic interest open-ended questionnaire to identified categories. | EH-5<br>Credibility and dependability of data analysis, research traceability and quality control are threatened by lack of consistent detail for all measures. Study is relevant to review question, but description of intervention is generalised.<br>**WoE**: ***medium-high*** **(9)** | No differences found between experimental and control group knowledge of historical content (pre- and post-test measure). Experimental group reported as scoring better than control group on the understanding of historical inquiry measure. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 12 | (Fontana et al., 2007)<br>USA | **59** Grade 10 and 11 students<br>**4** teachers | Two open-ended unit tests.<br>Two strategy use surveys.<br>One multiple choice post-test; Teacher and student satisfaction surveys.<br>Time sampling charts and tables. | Each measure scored by the researcher and at least one additional person. IRR calculated for each measure.<br>Rubrics used for scoring the various measures.<br>Sample answers for rater’s reference for scoring unit tests.<br>Students awarded 1 point for specific reference to any or all parts of the keyword strategy from intervention. | Two scorers for each measure. A third person was frequently enlisted to score or review the scoring and data for accuracy. Inter-rater reliability was calculated for each measure, and consensus was reached for each inconsistency.<br>Qualitative data were not analysed.<br>Time sampling to measure students’ behaviours.<br>Treatment fidelity checks. | EH-5<br>Study design does not account for intervention effect on classes receiving treatment first. Credibility and dependability of data analysis and quality control strong for quantitative, but absent on qualitative data. Study is relevant for the review question.<br>**WoE: *medium* (9)** | No immediate academic gains reported – intervention and control groups performed similarly; with exception of students for whom English was a second language – they scored significantly higher with intervention.<br>Intervention associated with higher levels of academic engagement, and preferred by students and teachers. |
| 13 | (Huijgen et al., 2018)<br>Netherlands | **131** Grade 10, 11 and 12 students | Pre- and post-tests. | Paired sample test to examine difference from pre- to post-test. Effect size calculated.<br>Multilevel analysis to identify the percentage of difference between experimental and control groups due to the intervention. | Authors and 4 experienced history teachers constructed 30 test items. 30 items piloted with 158 students. Authors devised another 8 items, and randomly assigned 19 items to pre- and 19 items to post-test to reduce *carryover effect.* 5 items in pre- and 5 items in post-test were found to threaten internal consistency and were deleted. Final pre- and post-test were evaluated by 2 expert history teachers and 2 educational measurement experts to ensure face and content validity.<br>Treatment fidelity checks. | EH-5<br>Credibility and dependability of study design, data collection and data analysis of quantitative items is *high.* However, the qualitative data is reported descriptively, with methods of analysis for these items absent. Study relevance for the review question is *high.*<br>**WoE: *high* (13)** | The results of the historical contextualisation test showed that students in the experimental condition demonstrated more progress in their ability to perform historical contextualisation compared to students in the control condition. A multilevel analysis indicated that the developed pedagogy had a medium effect on students’ ability to perform historical contextualisation. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 14 | (Nair and Muthiah, 2006)<br>Malaysia | **70** Grade 10 students<br>**2** teachers | Pre- and post-tests. Questionnaire. | Data processed using SPSS. *t*-test used to see the effects of the independent variables on every dependent variable. | Pilot study conducted for all instruments prior to the commencement of the study. Researcher observed the teaching of the experimental group. | EH-5<br>Credibility and dependability of study design and data collection threatened by lack of detail on intervention and data item questions. Quality control of data analysis threatened by single-coder bias. Study relevance for the review question is *low*.<br>**WoE:** ***medium*** **(8)** | Authors report experimental group performed significantly higher than control group for overall achievement in history. |
| 15 | (Nair and Narayanasamy, 2017)<br>Malaysia | **70** Grade 10 students<br>**2** teachers | Pre- and post-tests. Questionnaire. | Data from pre-test and questionnaire (prior to intervention) analysed using independent sample *t*-test. Data from post-test analysed using ANCOVA test (analysis of covariance). | Pilot study conducted for all instruments prior to the commencement of the study. | EH-5<br>Credibility and dependability of data collection, data analysis and quality control threatened by lack of detail on all items. Study relevance for the review question is *low*.<br>**WoE:** ***medium*** **(8)** | Authors report overall achievement in post-test shows the mean for the experimental group is higher than the control group. Results of the ANCOVA test indicate the experimental group performed significantly better than the control group. |

| | | | | | | |
|---|---|---|---|---|---|---|
| 16 | (Nokes et al., 2007)<br>USA | **246** Grade 11 students<br>**8** teachers | Observation.<br>Pre- and post-test for historical content.<br>Pre- and post-test for heuristics (historical epistemological knowledge and skills). | Historical content tests: ANCOVA with pre-test scores as covariate and post-test scores as dependent variable. Compared differences with post hoc multiple comparison tests on unadjusted means, effect size calculated using partial eta-squared, with follow-up multiple comparison test using Tukey's HSD.<br>Heuristics tests: ANCOVA used students' scores as the unit of analysis, with follow-up multiple comparison test using Tukey's HSD. | Observation: 2 observers used the instrument while attending 21% of lessons, inter-rater agreement 85.9%.<br>Historical content tests piloted with reliability calculated using Cronbach's alpha at .89. Content validity established by comparison to standardised NEAP and AP tests.<br>Heuristics tests author developed and piloted. Coding sheet and instructions supplied to 2 trained evaluators who both conducted blind assessment of 20% of pre- and post-tests with 78% agreement on heuristic scores | EH-5<br>Credibility and dependability of study design, data collection, data analysis and quality control were all *high*. Relevance for the review question was *high*.<br>**WoE: *high* (15)** | Historical content: Students who used multiple texts to study content scored significantly higher on post-test than all other groups. Students who used multiple texts to study heuristics scored significantly higher than students who used traditional texts to study heuristics. Higher performance on history content test attributed to using multiple texts.<br>Heuristics: students who used multiple texts to study heuristics scored significantly higher on sourcing post-test than did all other groups. No other significant differences.<br>Note: students used sourcing more than any other heuristic. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 17 | (Reisman, 2012)<br>USA | **236** Grade 11 students | 3 pre-test measures.<br>4 post-test measures. | One-way analysis of variance was used to test for differences between treatment and control groups in the reduced data set on the 3 pre-tests and on the composite covariate.<br>MANCOVA analysis for overall effect.<br>Follow-up univariate ANCOVA analysis for effect of 4 outcome measures: historical thinking, factual knowledge, reading comprehension, transfer. | 3 pre-test measures including California Standards Test (CST) and Gates–MacGinite Reading Test (GMRT-4)<br>4 post-test measures: Parallel GMRT-4; CST and New York State Regents Exam, US-selected items from both.<br>Curriculum fidelity checks.<br>Effects of treatment condition and schools examined with MANCOVA – 3 pre-test measure highly correlated, hence, principal component analysis conducted.<br>MCAR test to examine effect of missing student data. | EH-5<br>Credibility and dependability of study design, data collection, data analysis and quality control is *high*.<br>Study relevance for the review question is *medium* – intervention is not replicable.<br>**WoE: *high* (13)** | Author reports disparity in teacher curriculum fidelity scores, suggesting the intervention departed from teacher's normal instruction.<br>Significant overall effect on all outcome measures for both variables: treatment and school.<br>Significant effect for school on 3 outcome measures: historical thinking, factual knowledge and reading comprehension, but no effect on transfer of historical thinking. |
| 18 | (Stoel et al., 2015)<br>Netherlands | **43** Grade 11 students<br>**2** teachers | Pre- and post-tests. | Items in pre-test on which students scored very high leaving little room for improvement were excluded from the analysis. IRR recorded and Cronbach's alpha used where relevant. Essay questions were marked against a rubric using 7 criteria, and 2 raters independently scored essays, with discrepancies resolved through discussion. | Initial reliability analysis performed with students from different schools prior to implementation. Peer (historian and 2 history teachers) reviewed test items. Training of essay item raters and removal of 3 variables that reduced internal consistency. | EH-5<br>Credibility and dependability of study design, data collection, data analysis and quality control is *medium-high*.<br>Relevance for the review question: *high*<br>**WoE: *high* (13)** | Experimental and control conditions both improved causal strategies, with no detrimental effect on acquisition of content knowledge. Intervention enhanced knowledge of causal concepts and strategies. However, the application of concepts and strategies to essay task demonstrated smaller gains than expected. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 19 | (Van Straaten et al., 2019)<br>Netherlands | **1,022** Grade-10 to 12 students<br>**30** teachers | Pre- and post-tests. Questionnaires. | Multilevel analyses were conducted for each outcome measure. Model-fit was evaluated with the log-likelihood test and explained variance. Two types of effect sizes calculated: (1) standardised model-based effect sizes; and (2) effect sizes based on observed scores. Both model-based and observed effect sizes were standardised using Cohen's *d.* | Treatment fidelity checks. Use of a validated Relevance of History Measurement Scale (RHMS). | EH-5<br>Credibility and dependability of study design, data collection, data analysis and quality control threatened by absence of quality control for 2 measurement items. Intervention not replicable.<br>**WoE: *medium-high* (11)** | Authors report intervention had positive effect on students' perceptions of subject relevance; however, effects sizes were small. Experimental group considered lesson unit more valuable, and had less difficulty with applied pedagogical approach. Experimental group did not underperform in terms of historical knowledge acquisition. |
| 20 | (Wissinger and De La Paz, 2016)<br>USA | **151** Grade 6 and 7 students<br>**9** teachers | Pre- and post-tests. | Historical knowledge test: one-way MANCOVA. Box's M test, Shapiro–Wilk test, Wilks's lamba and ANCOVAs.<br>Reading Comprehension test: Author and 6 teachers scored using rubrics created for reading portion of PSSA.<br>Historical Thinking: analytic rubric to judge 4 elements – substantiation, perspective recognition, contextualisation, and rebuttal.<br>Writing quality: scored on 5 elements using PSSA persuasive writing rubric – focus, content, organisation, style, and conventions.<br>Essay length: scored on total number of words, irrespective of spelling | Treatment fidelity checks.<br>Historical knowledge and Reading comprehension test: author-teacher co-designed and reviewed by a separate teacher expert. Revised test used.<br>Historical thinking: Criterion validity established through bivariate correlations with the WIATT-III; 20 benchmark papers, 2 raters trained and then scored all pre- and post-tests.<br>Writing quality: Criterion validity established through bivariate correlations with the WIATT-III. 2 raters trained and scored all pre- and post-tests.<br>Essay length: Author counted total number of words in essays for pre- and post-tests, reliability checks were conducted. | EH-5<br>Credibility and dependability of study design, data collection, data analysis and quality control is *high.*<br>Relevance for the review question is *high.*<br>**WoE: *high* (14)** | Authors report the two argumentation schemes helped students to learn more historical content, and to demonstrate greater substantiation of evidence and more sophisticated rebuttals. The explicit teaching and discussion of responses to critical questions that accompany the argument from expert opinion facilitated students' growth of historical reasoning. |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 21 | (Wissinger et al., 2021)<br>USA | **608** Grade 4, 5 and 6 students<br>**11** teachers | Pre- and post-tests.<br>Maintenance post-test, 6 weeks after instruction ended. | Writing – analytic rubric from previous research. Essays scored on substantiation, perspective recognition, contextualisation and rebuttal on 3-point score, for a total of 12 points.<br>Reading comprehension measure was a non-equivalent dependent measure used to establish the Shadish et al. (2002) coherent pattern-matching principal. | Writing – 2 teachers trained, with one 'expert' rater given extra training. IRR established. Both raters independently scored all written response. IRR was 0.85, 0.89, 0.81, at pre-test, post-test and maintenance.<br>Reading comprehension – 2 trained research assistants, blind to study's purpose, scored the comprehension test, and 100% of the tests were checked for reliability. IRR was 1.0. | EH-5<br>Credibility and dependability of study design, data collection, data analysis and quality control is strong. Study relevance for the review question is *high*.<br>**WoE:** ***high*** **(15)** | Authors report intervention benefited students to a greater degree on all writing measures. Intervention had a moderate positive effect for substantiation, perspective, contextualisation and rebuttal. Authors report that with appropriate supports, academically diverse students in 4th to 6th grade can reason with primary source documents and write evidence-based historical arguments. |

## Nature of historical literacy pedagogical interventions in terms of being fit for purpose

Historical literacy pedagogies that are fit for purpose are those strategies that intentionally target discrete historical skills and knowledge acquisition. These strategies are typically scaffolded with explicit linkage to abstract skills required to think historically (for example, inferencing, interpretation, reasoning and judgement). All studies in this review with a high weight of evidence (WoE, 13–15) trialled pedagogies that guided students to think historically. Reisman (2012, #17) trialled a document-based approach offered by the Stanford History Education Group called Reading Like a Historian. In this approach, students used 'background knowledge to interrogate, and then reconcile, the historical accounts in multiple texts' (Reisman, 2012: 89), thus drawing on the skills of inferencing, reasoning and evaluation. Reisman reports gains in students' historical thinking. Similarly, Bertram et al. (2017, #4) found that their intervention groups demonstrated clearer understanding of the genre of oral sources, and were more likely to understand the constructed nature of historical recounts after engaging in 'oral history interviews in either active (live) or passive (video, text) ways' (Bertram et al., 2017: 453).

Studies in which Daniel R. Wissinger or Susan De La Paz have been involved (#6, 7, 8, 9, 20 and 21) typically trial discrete, often mnemonic, historical thinking and writing scaffolds.[2] For example: IREAD[3] for historical reading and annotations (see #8); I3C[4] for source analysis (see #21); H2W[5], STOP, DARE[6] or PROVE IT![7] for writing argumentative essays (see #8, 9, 20 and 21); a historical reasoning strategy graphic organiser[8] (see, #6 and 9); a Model of Domain Learning (MDL) framework for domain-specific content and pedagogical strategies (see #18); and heuristics[9] (see #7) to support disciplinary approaches to reading historical documents. Disciplinary approaches specific to history include perspective recognition, contextualisation of source material, corroboration of evidence and substantiation of argument. Authors of these studies report gains in students' capacity to make enhanced claims (#6, 7, 8 and 18), compelling rebuttals (#6, 7 and 20), persuasive arguments (#9) with greater substantiation of evidence (#20 and 21) and that identify perspective and the influence of context (#21).

Heuristics frequently feature in studies trialling pedagogical strategies to improve historical literacy (see also, #13 and 16). As Sam Wineburg (1999: 491) has argued, 'historical thinking, in its deepest forms, is neither a natural process nor something that springs automatically from psychological development'; hence, historical thinking requires learned and discipline-specific strategies. Nokes et al.'s (2007, #16) study intervenes with explicit instruction on the heuristics of contextualisation, providing students with opportunities to 'infer about the social and political context' (Nokes et al., 2007: 497) of sources through practice with multiple texts. They found in post-test data that students from the experimental group scored significantly higher on sourcing than all other groups. Success with historical contextualisation was also reported by Huijgen et al. (2018: #13), when the experimental group in the study demonstrated gains after participating in the pedagogical intervention that created cognitive incongruity to scaffold students' understanding of the importance of contextualisation and the dangers of presentism.

Not all pedagogical interventions trialled in this systematic review were fit for the purpose of improving students' historical literacy (#1, 2 and 3). The focus of Aidinopoulou and Sampson's (2017, #1) intervention was to compare the use of classroom time for student-centred activities between the flipped classroom model and the traditional classroom. While results reported gains for the flipped classroom model in historical thinking skills (HTS), the gain was rather in additional time for learning tasks 'such as collaborative activities and debates' that might 'cultivate' HTS (Aidinopoulou and Sampson, 2017: 242). Furthermore, a lack of data regarding what students did to demonstrate their understanding, analysis and interpretation of historical sources meant that we could not draw credible conclusions about the efficacy of the flipped classroom model for improvement in students' historical literacy skills. Similarly, the study by Ariës et al. (2015, #2) had insufficient information on the historical content of lesson activities and question types to enable us to make a credible judgement as to reported gains in historical reasoning. In this study (#2), identifying targeted discrete historical skills and knowledge acquisition was secondary to the meta-cognitive working memory training intervention. Historical literacy was also a secondary concern in Azor et al.'s (2020, #3) study, with their focus being firmly on the use of YouTube audiovisual documentaries for teaching history, and the effects on interest and achievement between genders. Moreover, there was insufficient detail provided on the pedagogy used with YouTube documentaries for us to make an assessment of the intervention's effectiveness (that is, fitness for purpose) for developing historical literacy.

## Steps taken in implementations of pedagogical interventions, and feasibility and replicability of historical literacy pedagogical interventions

The theory of change underpinning our systematic review of research into historical literacy education hypothesised that when a historical literacy pedagogical intervention was implemented in a primary or secondary school classroom, students would learn a discrete history thinking skill. Appraising the steps taken in the implementation of the intervention allows judgement as to the feasibility and replicability of the pedagogy. Approximately half of the studies in this review provided insufficient content and procedural detail on the pedagogical intervention to enable replication (#1, 2, 3, 5, 10, 11, 12, 14 and 15); however, if sufficient detail about the steps taken to implement the intervention was available, there remain a few studies that would not be feasible to replicate due to appropriateness or complexity of teaching and learning material required by the intervention. Studies falling into this category include #2, 3 and 12. Azor et al.'s (2020, #3) YouTube audiovisual documentaries intervention would not be appropriate to replicate, because the use of YouTube documentaries is not a historical literacy pedagogy in and of itself. Study #2 is not feasible to replicate because the complexity of the working memory training tool is highly specialised and atypical of the skillset of a history schoolteacher; the authors explicitly note this limitation to their study (Ariës et al., 2015). Finally, in Study #12 – as the authors (Fontana et al., 2007) observe – their intervention requires a sophisticated understanding of language to create their trialled mnemonic teaching and learning resource.

All studies in this review with a high weight of evidence (WoE, 13–15) provided sufficient detail for an expert practitioner to replicate the pedagogical intervention trialled. Most of these pedagogies followed a pattern of: (1) familiarisation with historical content; (2) explicit instruction; (3) expert modelling; (4) scaffolded learning activity; and (5) communication of learning (see #4, 6, 7, 8, 9, 13, 16, 17, 18, 20 and 21). The scaffolded learning activities targeted the development of historical epistemological knowledge and skills, and provided strategies for students to critique historical sources of information.

## Nature and extent of academic gains linked to interventions, and explicit or implicit moderating variables that may affect intervention results

Some claims of benefit to historical thinking from interventions trialled in this systematic review had insufficient data or quality controls to fully assess the credibility of conclusions (#1, 2, 3, 5, 10, 11, 12, 14, 15 and 19). Studies #1, 2, 5, 11, 14 and 15 reported improvements in students' historical epistemological knowledge and skills, with no detrimental effect on historical content knowledge acquisition. However, these studies are limited by the lack of identified historical literacy skills tested and reported on. While Studies #1, 2, 11, 14 and 15 identify reasoning or understanding historical sources, analysis or interpretation as the historical thinking skills developed during the intervention, there are no details provided as to the questions asked in the test instrument, nor is there sufficient description of the historical thinking lesson activities to enable external judgement as to the validity of their findings. Brugar (2016, #5) provides more detail in terms of historical thinking lesson activities; however, the qualitative data reported are at risk of single-coder bias. Further, statements of claim for students in the experimental condition showing the ability to draw inferences and make evaluations are not supported with reference to data items.

The extent of academic gains reported in Studies #10, 12 and 19 are also open to interrogation. Van Straaten et al. (2019, #19) reported gains in students' perceptions of subject relevance, with no underperformance in knowledge acquisition; however, their results are threatened by lack of quality controls reported for two measurement items (pedagogical questionnaire, and content knowledge post-test). de Groot-Reuvekamp et al. (2018, #10) reported gains in understanding historical time, with significant gains reported for Grade 5 participants; however, six teachers in the experimental condition spent on average an additional 24 minutes per week teaching history by providing students with both the intervention lesson and the traditional lesson. This treatment fidelity check was reported, but not factored into the analysis of intervention outcomes, therefore threatening the validity of assertions of academic gain due to the intervention pedagogy. Fontana et al. (2007, #12) reported significant gains for English as second language students for their mnemonic strategy intervention but 'no condition-specific performance differences' (Fontana et al., 2007: 352) overall in the post-test, suggesting that the intervention strategy has limited benefit for the development of historical literacy in general or mixed-ability history classrooms.

## Discussion

Our findings reported on the nature of historical literacy pedagogical interventions in terms of being fit for purpose, feasibility and replicability of pedagogy, academic gains linked to intervention, and explicit or implicit moderating variables that may affect intervention results. Findings of this review indicate that when a discrete historical skill or knowledge is targeted by a pedagogical intervention that utilises a scaffolded heuristic targeting explicit historical thinking skills, there is greater likelihood of positive outcomes for students learning historical literacy skills.

Studies in this review with a high weight of evidence (WoE, 13–15) demonstrated the most convincing and credible academic gains resulting from the pedagogy trialled. Studies in the 13–15 WoE category (#4, 6, 7, 8, 9, 13, 16, 17, 18, 20 and 21) were similar in terms of the instructional pattern adopted: teaching began with either teacher-directed or student familiarisation with historical content, sometimes combined with explicit teacher instruction; expert modelling followed; students next engaged in a scaffolded learning activity designed to improve their historical epistemological knowledge and skills; with the final step being a communication of findings drawn from results from the scaffolded learning activity. This finding is in line with Luís and Rapanta's (2020: 10) conclusion confirming the 'importance of a disciplinary approach to history teaching, one inspired by the use of empiricist historical thinking methods'. Our findings demonstrate how the more effective historical literacy pedagogical interventions work in terms of instructional patterning, and they are similar to other research findings noting a growing body of research demonstrating the effectiveness of strategies, such as the cognitive apprenticeship model, which includes 'explicit instruction, teacher modeling, opportunities for whole class and small group discussions, collaborative planning, and repeated practice with faded support [to] improve students' ability to produce written evidence-based historical argumentation' (Nokes and De La Paz, 2023: 357). These more effective pedagogies assist students to interpret, reason and explain historical evidence, analyse and synthesise evidence and historical argument, and provide a judgement expressed as an assessment of value or as an evaluation based on criteria.

The nature and appearance of scaffolded heuristics employed by the high weight of evidence studies reporting credible findings are targeted towards the deconstruction of historical sources, and thus assist students to identify elements such as: the context, audience or message of the source; the purpose of source creation and perspective represented; and techniques used to communicate the message, purpose and perspective of the historical source (see #4, 6, 7, 8, 9, 13, 16, 17, 18, 20 and 21). In addition, studies reporting credible findings include a scaffolded heuristic to develop students' communication of their epistemological knowledge and skills via reconstruction of historical source material through interpretation, reasoning, explanation, analysis and/or synthesis of evidence to construct historical arguments.

Our synthesis of studies in this systematic review has demonstrated how effective historical literacy pedagogical interventions work. The common feature of pedagogies trialled in the high weight of evidence studies is a scaffolded heuristic; hence, we have concluded that these interventions work because they provide students with explicit and discipline-specific step-by-step guidance on how to deconstruct and reconstruct historical sources and communicate results of findings. Nine out of the ten high weight of evidence studies were conducted in secondary schools (# 4, 6, 7, 8, 9, 13, 16, 17, 18 and 20), with only one study trialled in a primary school context, with Year 4, 5 and 6 students (#21). Based on the proclivity of studies trialled in secondary school contexts, we assume that high school provides the most appropriate context to trial pedagogical interventions designed to improve students' epistemological historical knowledge and skills; however, we acknowledge that the assumption is speculative.

A significant limitation of this systematic review is the inability to pinpoint which scaffolded heuristic is the most effective among the high weight of evidence studies. The considerable heterogeneity of the included studies, and the diversity in the comparisons being made by the primary studies, made it inappropriate to combine all included studies in a single meta-analysis. Given the mix of comparisons of different pedagogical interventions with different outcome variables, we are limited to describing key characteristics of interventions, rather than identifying the most impactful intervention for the purposes of improving historical literacy skills among primary and secondary school students. We further note that our study findings may be limited in their broader application, given that a high number of empirical studies included in this systematic review were conducted in the United States.

# Conclusions

The findings of this systematic review suggest that instruction using a discipline-specific scaffolded heuristic is beneficial to developing primary and secondary school students' historical literacy, especially in the area of epistemological knowledge and skills. Findings also indicate that when historical literacy development is of secondary purpose in a pedagogical intervention, the intervention is unlikely to effect a positive change in historical literacy skills. Findings demonstrate variability in the design of scaffolded heuristics, and inconsistency in measurement items across studies, thus making it problematic to compare the efficacy of individual studies to determine the most effective pedagogical approach to teaching historical literacy in primary and secondary school contexts. The field of research would be well served by a common measurement instrument to enable either the judgement or comparison of the efficacy of intervention trialled. Future research would also benefit from a broader international base of research sites, and detailed reporting of pedagogies, so that effective interventions could be replicated.

# Notes

[1] Krathwohl's (2002) Taxonomy of Educational Objectives has been referenced here because history curricula, irrespective of the documentation language or local regulatory stipulations, can typically be mapped to the revised taxonomy categories of: *A. Factual Knowledge*, *B. Conceptual Knowledge*, *C. Procedural Knowledge*, and *D. Metacognitive Knowledge*.

[2] We note the high number of studies originating in the United States, together with frequently cited researchers, Daniel R. Wissinger and Susan De La Paz. We assume that the proliferation of studies in this context is aided by the availability of funding for research into historical literacy, and supported by organisations such as the Stanford History Education Group.

[3] IREAD is a historical reading and annotation mnemonic to prompt students to: (1) ***IR*** Identify the author's purpose, Read each paragraph, and ask about the author's main ideas; (2) ***EA*** Evaluate the author's reliability, and Assess the influence of context; and (3) ***D*** Determine the quality of the author's facts and examples (De La Paz et al., 2017: 37).

[4] I3C is a historical reading strategy that prompts students to '(a) ***I***dentify the author's stance, (b) list ***3*** facts, ideas, or reasons supporting the author's stance, (c) ***C***heck for limitations in the author's argument by considering reliability issues and problems related to drawing inferences from perspectives in a single source' (Wissinger et al., 2021: 54).

[5] H2W, a 'How to Write Your Essay' graphic organiser (De La Paz et al., 2017: 38), includes essential components of historical argument and information signalling how to organise components of the composition.

[6] STOP is a mnemonic to prompt students, before writing, to '**S**uspend judgement, **T**ake a side, **O**rganise (select and number) idea, and **P**lan more as you write. DARE is a second mnemonic subroutine to prompt students to '**D**evelop a topic sentence, **A**dd supporting ideas, **R**eject an argument for the other side, and **E**nd with a conclusion' (De La Paz, 2005: 146).

[7] PROVE IT! is a historical writing strategy that prompts students to: '(a) ***P***rovide background information by describing the historical problem, (b) ***R***eport – or state – their interpretation, (c) ***O***ffer three reasons by including evidence from the documents, (d) ***V***oice the other side's interpretation, and (e) ***E***stablish a rebuttal. Students then had to consider (f) ***I***s the argument convincing. Finally, the students were asked to (g) ***T***otal up what they know by adding a sentence to conclude their essays' (Wissinger et al., 2021: 54–5).

[8] For example, the Historical Reasoning Strategy graphic organiser that requires students to '(i) Consider the author, (ii) Understand the source, (iii) Critique the source, and (iv) Create a more focused understanding' (De La Paz and Felton, 2010: 181).

[9] For example: Substantiation, Perspective recognition, and Contextualization (De La Paz et al., 2014: 251).

# Declarations and conflicts of interest

## Research ethics statement

The authors conducted the research reported in this article in accordance with Macquarie University research standards.

## Consent for publication statement

Not applicable to this article.

## Conflicts of interest statement

The authors declare no conflicts of interest with this work. All efforts to sufficiently anonymise the authors during peer review of this article have been made. The authors declare no further conflicts with this article.

# Appendix A. Search protocol

1. **Project Title**: Pedagogical interventions to develop historical literacy in primary and secondary school students: a systematic review
2. **Research question:** What is the relationship between pedagogical interventions and improved historical literacy in primary and secondary school children?
3. **Primary Outcomes:**

    3.1 Historical recount OR historical description OR historical narrative.

    3.2 Historical explanation OR historical interpretation OR historical judgement – in combination with accurate historical knowledge.

4. The Search:

    **Databases:** PsychInfo; ERIC; Academic Search Premier; Education Research Complete; Humanities International Complete
    **Search Terms:** Teach* Histor*; 'historical knowledge'; Historical recount; Historical description; Historical narrative; historical 'source analysis'; 'historical literacy'; School OR Primary School OR Elementary School OR Secondary School OR High School OR Freshman OR Sophomore OR Junior OR Senior OR Grades 1 through to Grade 12 OR Grammar School OR Grade School
    **Filters:** Date range 1990– 2021; Language – English; Journal Type – peer reviewed

5. **Screening:**

    **Inclusion Criteria:**
    **Inc.1** Study must take place with school-aged children (approx. 5–18 years old)
    **Inc.2** Instruction must be delivered in a school classroom and to the whole class group (either classroom teacher or guest teacher/lecturer)
    **Inc.3** Instruction occurs over a sustained period of time (e.g., one week or more, with a minimum of 3 sequential lessons)
    **Inc.4** Study must include ONE primary outcome
    **Inc.5** Classrooms must be in regular schools, this includes state schools and private schools
    **Inc.6** Empirical study
    **Inc.7** Includes the investigation of a pedagogical strategy.

    **Exclusion Criteria:**
    **Ex.1** Study took place at university (or another tertiary provider)
    **Ex.2** Instruction delivered outside a school classroom (e.g., in a museum) or to a small group (i.e., withdrawal group)
    **Ex.3** Instruction is provided as a one-off lesson/presentation
    **Ex.4** At least one primary outcome is not included
    **Ex.5** School or class caters for a specific population such as children with autism, poor hearing, learning difficulties etc.
    **Ex.6** Study is not empirical
    **Ex.7** Study does not investigate a pedagogical intervention
    **Ex.8** Case study.

    **Evidence Hierarchy for Included Studies:**
    **EH.1** Meta-analyses of RCTs
    **EH.2** Other meta-analyses (note if group comparison or single-case studies)
    **EH.3** Systematic reviews
    **EH.4** RCT (random control trial) **OR** CRT (Cluster randomised trials) – replicated or not

**EH.5** Quasi-experimental comparison group studies
**EH.6** Case study report
**EH.7** Expert reviews
**EH.8** Other school-based report.

## Appendix B. Critical Appraisal and Weight of Evidence (WoE) tool

<u>ARTICLE CITATION:</u>
<u>ASSESSOR:</u>

### A. Internal Methodological Coherence

**N.B**. A 'Can't tell' judgement must have an explanatory note provided.

| Criteria | | Judgement | | Notes |
|---|---|---|---|---|
| 1. | Did the study address a clearly articulated question(s) or issue (aim, objective or goal of study)? | Yes<br>No | 1<br>0 | |
| 2. | Is the study design appropriate to answer the question(s) or address the issue(s) (aim, objective or goal of study)? | Yes<br>No<br>Can't tell | 1<br>0<br>0 | |
| 3. | Is the Study Setting clearly described? | Yes<br>No | 1<br>0 | |
| 4. | Are the participants clearly described? | Yes<br>No | 1<br>0 | |
| 5. | Is the recruitment of participants clearly described? | Yes<br>No | 1<br>0 | |
| 6. | Is the sample design appropriate for the research focus? * | Yes<br>No<br>Can't tell | 1<br>0<br>*0* | *Identify sample design:* |
| 7. | Are the data collection procedures appropriate for the research focus? | Yes<br>No<br>Can't tell | 1<br>0<br>0 | |
| 8. | Are the procedures for data analysis reliable?<br>Check for use of quality control measures; for example, member checks, peer debriefing, attention to negative cases, independent analysis of data by more than one researcher, verbatim quotes, persistent observation, recursive design or constant reviewing of emergent themes and accurate representation of participants' voices. | Yes<br>No<br>Can't tell | 1<br>0<br>0 | |
| 9. | Is the research process traceable and clearly documented?<br>Check for the use of quality control measures, e.g., inclusion of sufficient data to assess credibility of conclusions, whether evidence can be inspected independently, peer review, calculation of inter-rater reliability agreement and triangulation | Yes<br>No | 1<br>0 | |

| | | | |
|---|---|---|---|
| 10. | Inclusion of enough information on researchers' orientations/background.<br>Check for the use of quality control measures, e.g., attention to the effects of the researcher during all steps of the research process, information on the researcher's background, education, perspective or relationship to study site. | Yes<br>No | 1<br>0 |

**Judgement Score for A. Internal Methodological Coherence /10**

**Overall internal methodological coherence rating:**

| High | High-Medium | Medium | Medium-Low | Low |
|---|---|---|---|---|
| 9–10 | 7–8 | 5–6 | 3–4 | 1–2 |

## B. Relevance of the study for the Review question

**Review question:** What is the relationship between pedagogical interventions and improved historical literacy in primary and secondary school children?

| Criteria | | Judgement | | Notes |
|---|---|---|---|---|
| 11. | Is the pedagogical strategy/ intervention clearly described?<br>Detailed = e.g., scaffolds identified and described, a lesson-by-lesson recount provided, lesson recount provides teaching sequences, pedagogical strategy/intervention could be replicated (by an expert teacher practitioner) based on the description provided<br>General = e.g., reference to a scaffold may be made but scaffold is not clearly described, an overview of a sequence of lessons may be provided (omitting lesson by lesson detail), lesson recount does NOT provide teaching sequences, pedagogical strategy/ intervention could NOT or most likely could not be replicated from the description provided. | Detailed<br>General<br>No | 2<br>1<br>0 | *Please provide explanation for judgement* |
| 12. | Is the historical literacy skill(s) targeted by the pedagogical strategy/intervention clearly defined? | Yes<br>No | 1<br>0 | |
| 13. | To which Primary Outcomes (PO) do the study findings report on?<br>PO(i) = Historical recount, description, or narrative<br>PO(ii) = Historical explanation, interpretation (incl. analysis), judgement (i.e., assessment, evaluation) – in combination with historical knowledge. | PO(ii)<br>PO(i) | 2<br>1 | |

**Judgement Score for B. Relevance of the study for the Review question /5**

**Overall Relevance of the study for the Review question rating:**

| High | Medium | Low |
|---|---|---|
| 5 | 3–4 | 1–2 |

**Combined Judgement Score for A. and B. /15**

**Overall rating for Weight of Evidence (WOE) scale:**

| High | High-Medium | Medium | Medium-Low | Low |
|---|---|---|---|---|
| 13–15 | 10–12 | 7–9 | 4–6 | 1–3 |

## References

Aidinopoulou, V. and Sampson, D.G. (2017) 'An action research study from implementing the flipped classroom model in primary school history teaching and learning'. *Educational Technology & Society*, 20 (1), 237–47.

Ariës, R.J., Groot, W. and Van den Brink, H.M. (2015) 'Improving reasoning skills in secondary history education by working memory training'. *British Educational Research Journal*, 41 (2), 210–28. [CrossRef]

Azor, R.O., Asogwa, U.D., Ogwu, E.N. and Apeh, A.A. (2020) 'YouTube audio-visual documentaries: Effect on Nigeria students' achievement and interest in history curriculum'. *The Journal of Educational Research*, 113 (5), 317–26. [CrossRef]

Bertram, C., Wagner, W. and Trautwein, U. (2017) 'Learning historical thinking with oral history interviews: A cluster randomized controlled intervention study of oral history interviews in history lessons'. *American Educational Research Journal*, 54 (3), 444–84. [CrossRef]

Brugar, K.A. (2016) 'Teaching social studies/history to elementary school students through a discipline-specific approach'. *Journal of Education*, 196 (2), 101–10. [CrossRef]

de Groot-Reuvekamp, M., Ros, A. and Van Boxtel, C. (2018) 'Improving elementary school students' understanding of historical time: Effects of teaching with "Timewise"'. *Theory and Research in Social Education*, 46 (1), 35–67. [CrossRef]

De La Paz, S. (2005) 'Effects of historical reasoning instruction and writing strategy mastery in culturally and academically diverse middle school classrooms'. *Journal of Educational Psychology*, 97 (2), 139–56. [CrossRef]

De La Paz, S. and Felton, M.K. (2010) 'Reading and writing from multiple source documents in history: Effects of strategy instruction with low to average high school writers'. *Contemporary Educational Psychology*, 35 (3), 174–92. [CrossRef]

De La Paz, S., Felton, M., Monte-Sano, C., Croninger, R., Jackson, C., Deogracias, J.S. and Hoffman, B.P. (2014) 'Developing historical reading and writing with adolescent readers: Effects on student learning'. *Theory and Research in Social Education*, 42 (2), 228–74. [CrossRef]

De La Paz, S., Monte-Sano, C., Felton, M., Croninger, R., Jackson, C. and Piantedosi, K.W. (2017) 'A historical writing apprenticeship for adolescents: Integrating disciplinary learning with cognitive strategies'. *Reading Research Quarterly*, 52 (1), 31–52. [CrossRef]

Del Favero, L., Boscolo, P., Vidotto, G. and Vicentini, M. (2007) 'Classroom discussion and individual problem-solving in the teaching of history: Do different instructional approaches affect interest in different ways?'. *Learning and Instruction*, 17 (6), 635–57. [CrossRef]

Fontana, J.L., Scruggs, T. and Mastropieri, M.A. (2007) 'Mnemonic strategy instruction in inclusive secondary social studies classes'. *Remedial and Special Education*, 28 (6), 345–55. [CrossRef]

Huijgen, T., Van de Grift, W., Van Boxtel, C. and Holthuis, P. (2018) 'Promoting historical contextualization: The development and testing of a pedagogy'. *Journal of Curriculum Studies*, 50 (3), 410–34. [CrossRef]

Krathwohl, D.R. (2002) 'A revision of Bloom's Taxonomy: An overview'. *Theory into Practice*, 41 (4), 212–18. [CrossRef]

Luís, R. and Rapanta, C. (2020) 'Towards (re-)defining historical reasoning competence: A review of theoretical and empirical research'. *Educational Research Review*, 31, 100336. [CrossRef]

Nair, S.M. and Muthiah, M. (2006) 'The effectiveness of using Needham's Five Phase Constructivist Model in the teaching of history'. *International Journal of Learning*, 12 (5), 311–22. [CrossRef]

Nair, S.M. and Narayanasamy, M. (2017) 'The effects of utilising the concept maps in teaching history'. *International Journal of Instruction*, 10 (3), 109–26. [CrossRef]

Nokes, J.D. and De La Paz, S. (2023) 'Historical argumentation: Watching historians and teaching youth'. *Written Communication*, 40 (2), 333–72. [CrossRef]

Nokes, J.D., Dole, J.A. and Hacker, D.J. (2007) 'Teaching high school students to use heuristics while reading historical texts'. *Journal of Educational Psychology*, 99 (3), 492–504. [CrossRef]

Pino, M. and Mortari, L. (2014) 'The inclusion of students with dyslexia in higher education: A systematic review using narrative synthesis'. *Dyslexia: An international journal of research and practice*, 20 (4), 346–69. [CrossRef] [PubMed]

Reisman, A. (2012) 'Reading like a historian: A document-based history curriculum intervention in urban high schools'. *Cognition and Instruction*, 30 (1), 86–112. [CrossRef]

Shadish, W.R., Cook, T.D. and Campbell, D.T. (2002) *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin.

Sterman, J., Naughton, G., Froude, E., Villeneuve, M., Beetham, K., Wyver, S. and Bundy, A. (2016) 'Outdoor play decisions by caregivers of children with disabilities: A systematic review of qualitative studies'. *Journal of Developmental and Physical Disabilities*, 28 (6), 931–57. [CrossRef]

Stoel, G., Van Drie, J. and Van Boxtel, C. (2015) 'Teaching towards historical expertise: Developing a pedagogy for fostering causal reasoning in history'. *Journal of Curriculum Studies*, 47 (1), 49–76. [CrossRef]

Toulmin, S. (1958) *The Uses of Argument*. Cambridge: Cambridge University Press.

Van Straaten, D., Wilschut, A., Oostdam, R. and Fukkink, R. (2019) 'Fostering students' appraisals of the relevance of history by comparing analogous cases of an enduring human issue: A quasi-experimental study'. *Cognition and Instruction*, 37 (4), 512–33. [CrossRef]

Weiss, C.H. (1997) 'Theory-based evaluation: Past, present, and future'. *New Directions for Evaluation*, 1997 (76), 41–55. [CrossRef]

Wineburg, S. (1999) 'Historical thinking and other unnatural acts'. *Phi Delta Kappan*, 80 (7), 488–99. [CrossRef]

Wissinger, D.R. and De La Paz, S. (2016) 'Effects of critical discussions on middle school students' written historical arguments'. *Journal of Educational Psychology*, 108 (1), 43–59. [CrossRef]

Wissinger, D.R., De La Paz, S. and Jackson, C. (2021) 'The effects of historical reading and writing strategy instruction with fourth- through sixth-grade students'. *Journal of Educational Psychology*, 113 (1), 49–67. [CrossRef]