Does explicit teaching of critical thinking improve critical thinking skills of English language learners in higher education? A critical review of causal evidence

*Nada El Soufi
Department of English Language and Literature
University of Balamand
Email: nada.soufi@balamand.edu.lb

Beng Huat See
School of Education
Durham University
Email: b.h.see@durham.ac.uk
Orcid id: 0000-0001-7500-379X
@beng_see

*Corresponding author

Abstract

This paper presents the results of a systematic review of international studies to establish whether explicit teaching of critical thinking is effective in enhancing the critical thinking skills of English language learners in higher education and to identify the most promising approaches. A search of 12 electronic databases supplemented by other sources yielded more than 1,794 studies. Only 36 met the pre-defined inclusion and exclusion criteria. A range of approaches were tested and almost all claimed to be effective, but only explicit instruction in general critical thinking skills was found to have the best evidence of effectiveness. However, because most of the studies were small-scale and/or methodologically flawed, the evidence is not strong enough to be conclusive. Evidence for the other approaches was even weaker. These findings suggest that research in this field is still rather immature and more large-scale, replicable robust studies are needed to advance the field.

Keywords: Critical thinking skills, systematic review, ESL/EFL, randomised controlled trial

1. Introduction

This paper presents the results of a systematic review of empirical evidence to establish the causal impact of explicit teaching critical thinking skills for English language learners (those

for whom English is not their first language) in higher education, and to identify the most promising strategies. In order to establish causality, only studies using an experimental or quasi-experimental designs were considered in this review. Findings from this review will therefore be relevant to higher education educators and instructors in second language classrooms. It will provide evidence on the best approach to use that facilitates the teaching and learning of critical thinking skills.

2. Background

Traditionally the role of universities has been to develop independent and critical thinkers (Mitchell et al., 2003; Halpern, 2014) able to judge the trustworthiness of evidence and distinguish facts from opinions (Renaud & Murray, 2008). The increasing marketization of higher education and the focus on the university as an economic enterprise, however, has somewhat turned the focus of the university away from this traditional role (See, 2016). Nevertheless, there is no denying that the ability to think critically is even more relevant today with the proliferation of information from all sources such as social media, and the recent phenomenon of "fake news". More than ever before young people need to be able to discriminate facts from opinions, evaluate and judge the credibility of evidence. An effective strategy to foster such skills is through the development of critical thinking skills (Driver et al., 2000; Sadler, 2006). Critical thinking is also increasingly sought after in the workplace. An examination of 4.2 million job advertisements in Australia between 2012 and 2015 reveals that demand for employees who have critical thinking skills has risen by 158% (Foundation of Young Australians, 2016).

The role of education in fostering critical thinking in students has been stressed since the time of Dewey (1910). Contemporary thinkers and educators (Pithers & Soden, 2000; Davies, 2003; Marin & Halpern, 2011; Moore, 2011) hold similar views. In Europe, the reform in science education in 2011 made the teaching of critical thinking a main aim of undergraduate teaching (Eurydice, 2011). Similarly in the US critical thinking was identified as one of the key learning outcomes for all undergraduates (Association of American Colleges and Universities, 2004, 2015). Academics in Australia also concur that critical thinking is an essential skill in higher education even though they have different understanding of what critical thinking is (Moore 2014).

However, despite the emphasis on critical thinking in higher education there is little evidence that such skills are taught in an explicit and systematic way at undergraduate level (Coil et al., 2010). Research in the US found that many university graduates lack the skills to distinguish facts from opinion or make clear written argument or objectively review conflicting reports (Arum & Roksa, 2011; Shim & Walczak, 2012). See's (2016) study of UK universities found that the first year course modules of most disciplines do not explicitly teach critical thinking, and that students were less likely to be taught critical thinking at university than at school. Most first year courses emphasised the dissemination and recall of factual knowledge. Schafersman (1991) also noted that it was the emphasis on the acquisition of basic knowledge that had led to the neglect of the role of university as an institution for developing critical thinkers. A similar concern was expressed about the heavy reliance on rote memorization of decontextualized information in higher education in developing countries (Richmond, 2007). This is particularly so in some countries where it is considered rude to question authorities, and where argumentation and questioning evidence is not encouraged. In the last three decades commentators have explicitly expressed the inability of higher education students to use higher-order thinking skills (e.g. Norris, 1985; Gimenez, 1989; Halpern, 1993; Paul, Elder & Bartell, 1997; Blackmore, 2001; Pally, 2001; Paul, 2004; Davies, 2011; Marin & Halpern, 2011).

In his article The State of Critical Thinking Today, Richard Paul (2004) identified three main obstacles to acquisition of critical thinking in higher education. First, universities are not aware of their lack of substantive concept of critical thinking. Second, they thought they knew what critical thinking is and are already teaching students it. Third, lectures, rote memorisation and short-term learning habits are the norm in higher education. Some believe that critical thinking is a single-subject discipline and thus teach critical thinking as logic or study skills. As a result, you have a situation where lecturers expect students to be able to analyse complex concepts, but have no idea how to teach it. They expect intellectual standards from their students, but do not have a clear idea of what is considered an intellectual standard or how to formalise it.

It was the general dissatisfaction with students' inability to reason well that started the critical thinking movement in the 1980s (Facione 1990). Known as the Delphi Project, the movement, led by Peter Facione and sponsored by the American Philosophical Association, brought together a body of international philosophers, scientists, and educators to define critical thinking and to give recommendation on critical thinking instruction and assessment. They

defined critical thinking as "purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, or contextual considerations upon which that judgment is based" (Facione, 1990, p. 2).

Since then the term 'critical thinking' has been defined variously as critical reasoning, argumentation, critical evaluation and higher-order thinking. For the purpose of this paper, we define critical thinking to include Toulmin's model of argumentation. Critical thinking is about the ability to make arguments and Toulmin's definition of argumentation involves critical thinking skills. Andrews (2015) argues that critical thinking and argumentation are closely related and both have implications for teaching and learning in higher education. For the purpose of this paper, we define critical thinking as the ability to understand assumptions, make claims that are supported by evidence and make conclusions that are warranted by the evidence presented.

While many educators agree that critical thinking is an important skill, not all agree on the best approach to teaching it. Dissent among educators lies in whether critical thinking is a generic set of skills that can transfer across domains and that can be taught independent of subject or whether it is domain-specific (McPeck, 1984; Bailin et al., 1999; Moore, 2014) and should be taught explicitly.

Some commentators argue that critical thinking is a cultural practice and cannot be easily taught (e.g. Atkinson, 1997; Ramanathan & Kaplan, 1996). Such claim is concerning as it means that developing critical thinking in non-native English language learners is almost impossible. In cultures that attach importance to conformity and discourage independent thinking, fostering critical thinking is all the more relevant. Unfortunately, in such countries language teachers are often more concerned with language accuracy than critical appraisal of texts. English language classes in these countries often involve students reading a text and answering comprehension questions. Rarely are they asked to evaluate the text, or judge the credibility of the information. In many cases, the materials used in the language classroom do not encourage students to think critically. It is the aim of this review to determine whether critical thinking can indeed be taught in the language classroom and if so, what is the most effective approach to teaching it.

Previous systematic reviews and meta-analyses tended to focus on different instructional approaches to instruction in critical thinking (Abrami et al., 2008; Tiruneh, Verburgh, & Elen, 2014), relationship between research design, type of assessment and effect size (Abrami et al., 2008; Behar-Horenstein & Niu, 2011), argumentation skills in various disciplines (Torgerson, Andrews, Robinson, & See, 2006), and different approaches to instruction in critical thinking in all subjects and overall university experience (McMillan, 1987; Ten Dam & Volman, 2004; Niu, Behar-Horenstein & Garvan, 2013; Abrami et al., 2015; Huber & Kuncel, 2016). As far as we know, there have been no reviews that examined the teaching of critical thinking skills to English language learners (students whose first language is not English) in higher education.

The present systematic review therefore fills this gap. It is the first to focus on the teaching of critical thinking skills in the English language classroom at the university level where thinking in the target language is required.

3. Research questions

This systematic review focuses on instruction in critical thinking in the English language learning classroom (also known as ESL/EFL) in higher education. ESL stands for English as Second Language, and EFL stands for English as a Foreign Language. The main research questions are:

- Is explicit instruction in critical thinking feasible in the English language classroom?
- What are the most promising approaches for teaching critical thinking skills to English language learners in higher education?
- What are the least effective approaches to teaching critical thinking skills?
- What are the barriers to teaching critical thinking to English language learners in higher education?

The main aim of the present systematic review is to identify the most effective approach to instruction in critical thinking in the language classroom. Therefore, only studies that can establish a causal relationship between instruction in critical thinking and the level of critical thinking are considered. For this reason we included only studies that use experimental and quasi-experimental designs (e.g. randomized controlled trial, regression discontinuity, difference-in-difference and studies using matched comparison, instrumental variables or propensity score matching). Cross-sectional correlational studies, while useful, cannot

determine causation since they cannot control for other unobservable confounding factors. These studies would therefore not be able to add to the evidence base. Experimental studies, on the other hand, form the best warrant to confirm causal conclusions (Cook & Shadish, 1994; Cook, 2002; Shadish, Cook, & Campbell, 2002; Robson, 2014).

4. Methods

The review involved a series of steps, beginning with locating the literature and screening for relevance. Each included study was then data extracted and quality appraised so as to judge the trustworthiness of the evidence. The studies were then synthesized to identify the most promising approaches. For this reason only empirical research using experimental or quasi-experimental designs were considered in the analysis.

4.1 Identification of studies

The review began with a comprehensive search in twelve relevant electronic databases. These were British Periodicals, Social Science Database, ERIC, International Bibliography of the Social Science, Periodicals Archive Online, ProQuest Dissertations & Theses, ProQuest Dissertations & Theses Global, Education Database, PsychINFO, British Education Index, Web of Science, and JSTOR. A further search using Google and Google Scholar was also conducted to identify grey literature and unpublished literature to reduce the possibility of publication bias. Relevant studies in the reference list of identified studies were also followed up.

The search was limited to those reported or published in the English language from 1990 to November 2018. We were specifically looking for studies related to the teaching of ESL/EFL courses for students above the age of 16. Studies were excluded if they were about the use of technology, English for Academic purpose, grammar, phonology, literature, gifted students or students with disabilities, and metacognition. Studies that only dealt with the assessment of critical thinking without an intervention were also excluded, as were studies that simply described critical thinking approaches. As the purpose of the review was to identify teaching approaches that enhance critical thinking (a causal question) only studies that used experimental and quasi-experimental designs were considered. Correlational, observational studies, opinion or thought pieces, theoretical/philosophical views on critical thinking and narrative accounts of the researcher's experience would not be relevant to the research questions as they cannot determine causation.

A search of the databases and internet search engines was conducted using the following keywords and their synonyms:

("critical thinking" OR "critical reasoning" OR "higher-order thinking" OR "rational thinking" OR "analytical thinking" OR "cognitive skills" OR argument* OR debate* OR "thinking skills" OR criticality)
AND
("language teaching" OR "language learning" OR "foreign language" OR L2 OR L1 OR "second language" OR ESL OR EFL OR "target language" OR "English language" OR "language skills")
AND
(intervention OR experiment* OR "quasi-experiment*" OR "difference in differences" OR study OR "randomized controlled trial" OR "regression discontinuity" OR factorial OR "controlled study")

The search was run a number of times with different search options and limiters to make sure no relevant studies have been missed, and adjusted to suit the idiosyncrasies of the different databases.

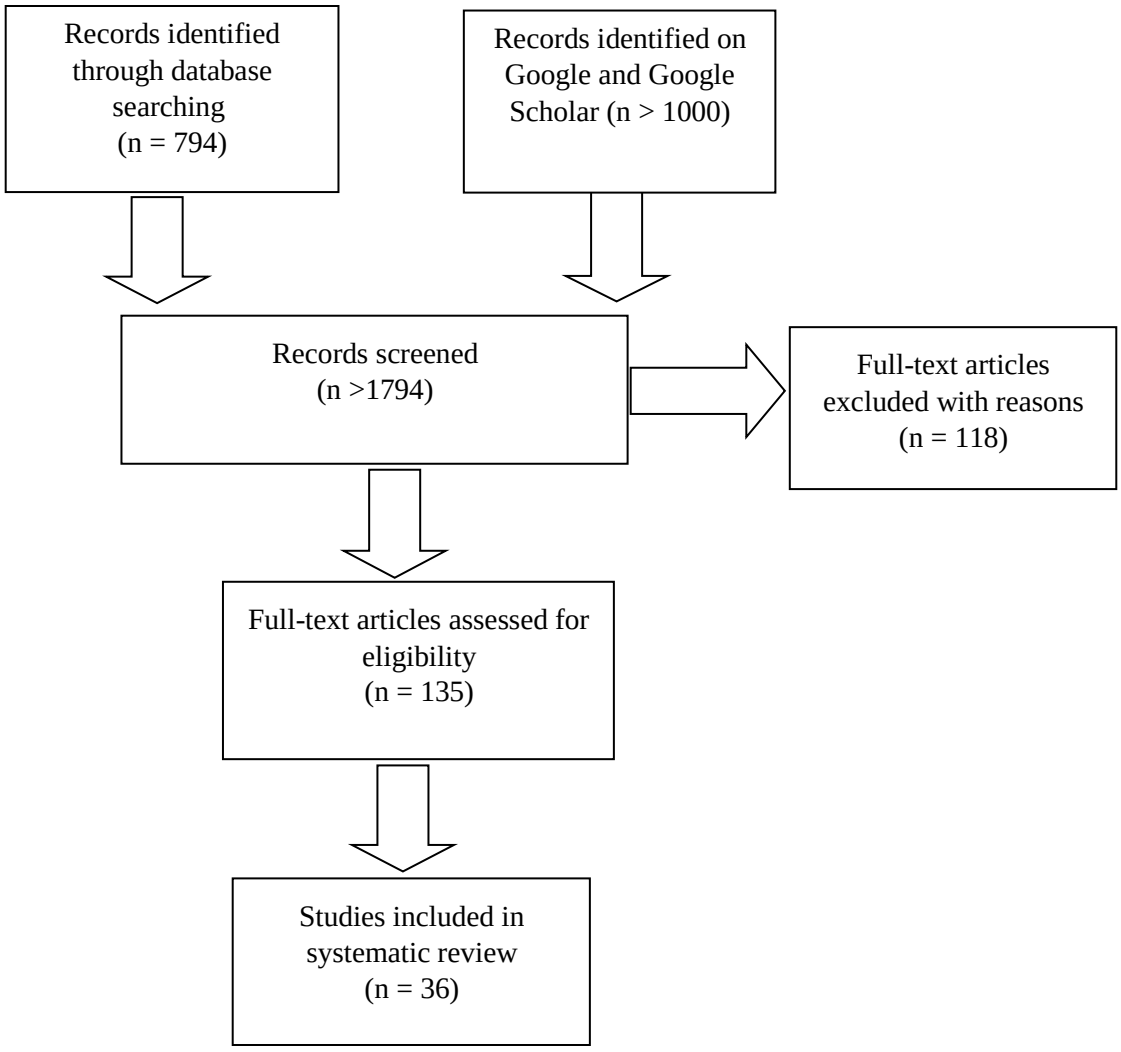
4.2 Cleaning the data

The database search yielded 794 results and an additional 1,000 from handsearching of google and google scholar. These were then imported to Zotero, and the titles and abstracts were screened for duplicates and relevance. A large proportion of these were duplicates. This is not surprising since there is a huge overlap of journals in the different databases. After screening only 135 were judged to be relevant and not duplicates. Of these, 118 were further excluded for various reasons like not being done in a language classroom, not being primary research, dealing with students below the age of 16, dealing with technology, or dealing with students with special needs. Only 36 studies met the inclusion criteria and were data extracted. The PRISMA flowchart (Moher et al., 2009) shows the number of records identified and the number of included and excluded studies at each stage.

Data extraction involved extracting information about all aspects of the research design which include matters pertaining to the sampling strategy, the sample size, allocation to groups, the instrument used to assess the outcome measure, and the attrition rate.

Flowchart 1

PRISMA flowchart



4.3 Appraisal of the quality of studies

An important aspect of the review was the quality assessment of individual studies to determine how much confidence could be placed on the findings. This is necessary to ensure that the evidence is trustworthy, and this is where our review differs from most systematic reviews which are only concerned with the results and do not discriminate between poor quality evidence and credible evidence. We do not accept the source of any publication or the status of its author or funder as any guarantee of research quality. Instead we judge the quality of evidence for each of the 36 included studies applying a quality assessment tool, known as the "Sieve". The Sieve, designed by Gorard (2014) (see Table 1), was specifically designed for educational interventions. It is an objective and structured way to appraise

studies. The “Sieve” is a tool for assessing the security of research findings and has been adopted by the Education Endowment Foundation in their padlock ratings for their evaluations (https://educationendowmentfoundation.org.uk/public/files/Evaluation/Carrying_out_a_Peer_Review/2016_Classifying_the_security_of_EEF_findings.pdf).

Table 1: A "Sieve" to judge the trustworthiness of experimental research

Design	Scale	Dropout	Outcomes	Fidelity	Validity	Rating
Fair design for comparison	Large number of cases per comparison group	Minimal attrition, no evidence of impact on findings	Standardised pre-specified independent outcome	Clear intervention, uniform delivery	No evidence of diffusion or other threat	4★
Balanced comparison	Medium number of cases per comparison group	Some initial imbalance or attrition	Pre-specified outcome, not standardised or not independent	Clear intervention, unintended variation in delivery	Little evidence of diffusion or other threat	3★
Matched comparison	Small number of cases per comparison group	Initial imbalance or moderate attrition	Not pre-specified but valid outcome	Unclear intervention, with variation in delivery	Evidence of experimenter effect, diffusion or other threat	2★
Comparison with poor or no equivalence	Very small number of cases per comparison group	Substantial imbalance and/or high attrition	Outcome with issues of validity or appropriateness	Poorly specified intervention	Strong indication of experimenter effect, diffusion or other threat	1★
No report of comparator	A trivial scale of study, or N unclear	Attrition not reported or too high for any comparison	Too many outcomes, weak measures, or poor reliability	No clearly defined intervention	No consideration of threats to validity	0

Source: Gorard, 2014

Each study was given a star rating based on six criteria: the design (e.g. whether it is an RCT with random assignment of cases, if there is a comparator group, or matched comparison),

scale of the study (sample size), level of attrition, how outcomes are measured (e.g. standardised tests, self-report, intervention-related or developer constructed instruments), fidelity and threats to validity. Ratings ranged from 5* to 0. Five star studies are the most secure, meaning that the evidence is most reliable or trustworthy. The "Sieve" rating reads from left to right and from top to bottom. For example, studies with a fair comparison as in an RCT would start with a 5* rating, and moving to the next column, if it has a very small number of cases in each group it would drop to 1*

Therefore, small studies involving randomising two classes to treatment or control condition would be given a lower rating despite being considered an RCT since the two classes may be different in terms of student and teacher characteristics. Therefore any impact may be due to these differences and cannot be solely attributed to the intervention. A cluster, an intact unit, or a higher order unit, as Shadish, Cook and Campbell (2002) explain, is very common in educational settings. These could be classes, schools or districts. Randomizing clusters is not the same as randomizing individuals as individuals within clusters might have common inherent qualities. Shadish, Cook, and Campbell (2002) warn that the individual participants within the higher order unit, cannot be treated as independent of each other as they are exposed to the same influences other than the treatment such as the teacher. Therefore, where classes are randomized, there needs to be a big number of clusters (or classes). Each cluster or class is therefore considered a case. If two classes are randomized then the number of cases is only two regardless of the number of students there may be in each class.

Studies with no comparators would immediately be given a low rating because without a comparison, it is not possible to attribute any changes to the intervention or programme.

Threats to the internal validity of any study would also involve an examination of the number of counterfactual cases needed to disturb the findings (NNTD). NNTD is a method of assessing whether the level of attrition (missing cases and missing data) would have altered the results. Missing cases and missing data are seldom random. Those that drop out of a trial or did not answer certain questions are likely to be different to those who did. By not considering missing cases, it is likely to overestimate the effect. Therefore, it is important to consider attrition. As Gorard, See, and Siddiqui (2017) explain, calculation of the number needed to disturb the finding can reveal whether the study would result in completely different findings if more cases were added to the smaller group. Calculation that yields a number bigger than the number of

missing cases means that it takes many more cases in order to have different results and therefore the findings are considered to be secure. NNTD is calculated by multiplying the effect size by the number of participants in the smaller cell. The bigger the number is, the more secure or stable is the finding. If the number of missing cases is small in comparison with NNTD, the finding can be safely considered not to be the result of mere chance due to attrition.

A number of studies in the review did not calculate effect size so it was hard to tell whether there was positive impact or not. We do not simply accept the authors' claims of effectiveness. Where effect size was not calculated by the original author(s), we did the calculation using the data presented. Effect size is "a way of quantifying the size of the difference between two groups" (Coe, 2002, p. 1). The effect size used here is the Hedge's g effect size, which is calculated by subtracting the post-test mean of the experimental group from the post-test mean of the control and dividing it by a pooled standard deviation of both means.

To ensure inter-rater reliability, the ratings were completed by two raters who both rated the studies individually and then compared the ratings. Where there was a disparity, ratings were explained and an agreement was reached.

4.4 Synthesis

Each of the 36 studies was allocated a star rating indicating the strength of the evidence. In addition, the NNTD was also calculated to establish the security of the findings where missing data may skew the results. A number of studies in this review did not report attrition nor provide sufficient data for the calculation of effect size. Inadequate or shoddy reporting is a reflection of poor research. These were therefore rated low in terms of quality. We cannot assume that because data is not reported clearly or in full, we should just accept its findings.

For the purpose of this paper we will discuss the higher rated studies in more detail. Those with a 0 rating will not be extensively discussed as their findings do not contribute to the evidence that will answer the research questions.

Once rated the studies were synthesized according to the approaches or strategies used in the teaching of critical thinking. The outcomes of each of the approach (i.e. positive or negative effects), the ratings for each of the studies are presented in a table. Approaches with the most number of positive studies do not necessarily mean that they are the most effective.

Consideration has to be taken of the quality (star ratings). This means that approaches with the most number of high quality ratings showing positive effects are considered most promising and those with the lowest number of high quality studies would be regarded as less promising.

(See Appendix for more details of the quality assessment of the included studies)

5. Findings

The 36 studies examined in this review have used a variety of approaches for instruction in critical thinking. The most common instructional approaches found in this review concerns teaching general critical thinking skills (n = 13 studies), followed by the use of literary and narrative texts (n = 6) and assessment techniques (n = 5) like peer-review, teacher evaluation, and self-evaluation. Other approaches include the use of debates, brainstorming techniques, journal writing, scaffolding, and active learning strategies. Almost all studies claimed positive effects, but most were given very low ratings. For this reason we think it is necessary to discuss some of these weaker studies to justify our ratings.

5.1 Most promising approaches to teaching critical thinking in higher education

This section describes the most promising approach to teaching critical thinking skills. No studies were found to be of good quality or even of medium quality due to serious flaws in their design. Therefore, there is no strong evidence that any instructional approach for teaching critical thinking skills works. However, instruction in general critical thinking skills looks potentially promising as it has been examined by a bigger number of studies than other approaches and all the higher quality studies reported positive effects. In addition, the approach itself seems plausible enough to maybe lead to some growth in critical thinking.

5.1.1 General critical thinking skills

Instruction in general critical thinking skills involves training students to define arguments, evaluate reliability of sources, identify fallacies and assumptions, use inductive and deductive logic, synthesize information, make inferences, etc.

This approach has been evaluated in the most number of studies, and all, but two reported positive effects (Table 2). Although two studies reported negative effects **(Zelizer, 2013; Manning, 1997)**, their evidence is very weak. Zelizer's (2013) study, for example, did not evaluate the effectiveness of critical thinking instruction. Instead it compared two different

approaches to teaching critical thinking (mixed instructional approach with an immersion approach). Also some of the lessons were taught by the same instructor, which might have resulted in diffusion of treatment. Participants who did not complete the post-test were excluded from analysis. This meant that the results are unreliable as participants who dropped out from the study could be different from those who complied. Manning (1997) compared two groups of very different students on campuses 30 miles apart. The experimental students were mature students and many with family and work responsibilities. The comparison groups were therefore not equivalent to begin with. We can therefore safely discount their evidence.

Table 2: Quality and impact summary: Studies focused on instruction in general critical thinking skills (N = 13)

Author(s) + Year	Smallest cell	Attrition	Effect size	NNTD	Quality
Salmani Nodoushan (2016)	Not specified	1.34% (12)	0.01 (calculated by reviewers) – essay 35.6 (calculated by reviewers) – The Cornell Critical Thinking Test, Form Z	-	2*
Gomez (2010)	40	18% (15)	0.08 (calculated by reviewers)	3	2*
Mazer, Hunt, & Kuznekoff (2007)	155	Not reported	0.34 (calculated by reviewers)	53	2*
McCarthy-Tucker (1995)	57	38.8% (120)	0.33 (calculated by reviewers)	19	2*
Ruff (2005)	19	Not reported	Not enough data provided	-	1*

Davidson & Dunham (1997)	17	14% (5)	Not enough data provided	-	1*
Zelizer (2013)	79	8% (14)	-0.08 (calculated by reviewers)	0	1*
Dong (2017)	22	Not reported	1.89 (calculated by reviewers)	-	1*
Akbari, Seifoori, & Ahour (2017)	25	Not reported	Not enough data provided	-	1*
Moore (1995)	Not applicable	Not reported	Not applicable*	-	0
Turuk (2011)	9	47% (7)	Not enough data provided	-	0
Manning (1997)	15	Not reported	-0.89 (calculated by reviewers)	-	0
Chason, Loyet, Sorenson, & Stoops (2017)	Not applicable	Not reported	Not applicable*	-	0

* single-group design – no comparison of gainscores

Of the eleven studies that reported positive effects, four were given a rating of 2* - the highest rating in this review.

Four studies reporting a positive effect of the generic approach to critical thinking were rated 2*. The first was a randomized controlled trial involving 894 students from different universities in Iran (**Salmani Nodoushan, 2016)**. Students were randomly assigned to either treatment or control group in each of the four language proficiency groups (limited English proficiency, lower intermediate, upper intermediate and advanced). Only 12 students dropped out. Experimental students were offered a 3-week workshop in their mother tongue, Persian, to raise participants' awareness of critical thinking strategies and in particular fallacious argumentation. The rationale behind using the mother tongue of the participants was to avoid the extra support that the experimental group would get in writing that the control group would

not receive, which might affect the performance of the experimental group in writing. Students were given the post-test after a two-week interval. The Cornell Critical Thinking Test was given in Persian. This is the only study in which the use of students' native language is justified as the researcher's aim is to investigate whether L1 mediated learning that aims at enhancing students' critical thinking skills would improve their argumentative writing

This was rated 4* initially for its scale and design, but dropped a star to 2* because the intervention materials were identical to the items used in the Cornell Test. Effectively, the researchers were teaching to the test. Another problem with the study is that raters of the essays were not blinded, which might have skewed the results in favour of a particular group due to teacher expectation. The effect size of the Cornell Critical Thinking Test would be +35.6, which is extremely unlikely – something that has never been seen before in any trials. This immediately puts suspicion on the reliability of the findings. As Bob Slavin says: "the chances of finding effect sizes of more than +1.00 are the same as the chances of finding a 10-foot man", assuming that the test was not a test of the intervention materials which the control group had no access to (Slavin, 2018)

The second study (**Gomez 2010)** involved 86 first year university students who were individually randomized to receive the intervention or business-as-usual. Students in the control groups were taught with emphasis on basic reading comprehension skills and adhered to the activities that are in the textbook whereas students in the experimental groups had more expansion activities that included analysis, application, evaluation, and synthesis of the material. Outcomes were measured using the translated version of the standardized California Critical Thinking Skills Test (CCTST). A small positive effect (ES = +0.08) was observed after one semester lasting 15 weeks. The small effects could be because the test was in Spanish while the instruction was in English. This might have worked against the students as students might have become used to thinking in a particular language in the classroom, so they could not transfer what they had learned using a particular language to the test which is administered in another language. This problem of transfer from one language to another is particularly problematic for students who are novice critical thinkers. Although this was a well-designed study and could have been a 4*, the poor choice of instrument, the relatively high level of attrition (18%) coupled with the small sample size meant that the highest rating could only be 2*. The NNTD is only 3 as opposed to an attrition rate of 15 participants. The evidence is therefore weak, but the results are promising.

The third study by **Mazer, Hunt, and Kuznekoff (2007)**, was also a cluster randomised controlled trial where 18 clusters of 324 university students ranging from age 18 to 26 were randomly assigned to treatment conditions. Experimental students were explicitly taught critical thinking skills. The control students followed the routine course structure. Outcomes were measured using a bespoke critical thinking test developed by the researchers. Experimental students made bigger gains than control students. This study could be rated more highly but because the outcome was measured using a researcher-developed test, it is possible that the teacher/researcher could have taught to the test, or the test could be intervention-related. Attrition was also not reported. All this lowers the credibility of the study and hence the 2*.

Another cluster randomized trial with a 2* rating also reported positive effects. In this study **McCarthy-Tucker (1995**) allocated 9 clusters of students (N = 309) to two groups to examine whether instruction in formal logic can improve students' critical thinking in English and maths. Outcomes were measured using the Raven's Standard Progressive Matrices (RSPM) and Test of Logical Thinking (TOLT) and the Content-Specific Test of Logic (CSTL). Although the study design is strong, the high attrition of nearly 40% meant that the findings are no longer reliable. The study was therefore given a 2* rating. Only the scores of students who took the pre-test and post-test and attended at least 85% of the instruction were included in the analysis. An intention-to-treat analysis and a compliance analysis could have been conducted to see if those who dropped out differed in any way from those who did not. The NNTD is 19 compared with 120 missing cases. Therefore, the findings have to be considered with caution.

Two other positive studies were rated 1* as they were weaker in design being quasi-experiments. The first study **(Davidson & Dunham 1997)** was a two-group post-test only design. The study, spanning over a year, compared 17 students enrolled in an intensive academic programme with a group of 19 volunteers who served as the control. Experimental students received training in critical thinking skills. Outcomes were measured using the Ennis-Weir Critical Thinking Essay Test. Results showed that the experimental students did better than the control but without a pre-test it was difficult to say which group had made bigger progress. It is possible that the experimental students who signed up for the course may have higher scores to begin with. But this was not measured. The lack of data, unclear reporting

about the allocation process and the very small sample size meant that the reported findings have to be treated with caution and hence the 1*.

In another quasi-experimental study, **Ruff (2005)** compared students enrolled in a transitions course in which critical thinking was taught (n = 20) with students who were enrolled in the same course but did not receive instruction in critical thinking (n = 19). The groups were not randomly allocated. Different textbooks were used for the two groups but the course was taught by the same teacher. There is therefore a possibility of diffusion. Experimental students were given activities that involved analysis, interpretation, evaluation, and synthesis while the control group did not have any exposure to critical thinking skills. Students were tested before and after the intervention using the California Critical Thinking Skills Test (CCTST) and the California Critical Thinking Dispositions Inventory (CCTDI). These are standardized tests of critical thinking. Although the author reported positive effects, no effect size was calculated and there was no enough data reported for any effect size to be calculated. There was also no report of attrition. This study was therefore rated a 1*.

Dong (2017) examined the effect of integrating a critical thinking approach based on Paul and Elder's (2001) CT model in a writing course on students' level in critical thinking. The researcher randomized two intact classes. The original writing teacher taught the control group while the researcher taught the experimental group. The major weakness in this study is that because the two classes were taught by different instructor, it was not possible to control for teacher effect. Therefore, we cannot confidently attribute any gain in critical thinking in the experimental group to the intervention. Randomizing two intact classes means that the number of cases is two. Students within a cluster could be exposed to the same influences (for example teacher effect) and there could also be unobservable differences between the two groups. Although the essays were graded by two teachers, teachers were not blinded, which could have biased the results. The study was therefore given a rating of 1*.

Akbari, Seifoori, and Ahour (2017)'s study also randomised two intact classes, thus reducing the number of cases to two. The researchers ensured balance between the two group in terms of language proficiency by administering the TOEFL test and then choosing 50 students out of 60 who scored ±1 standard deviation of the mean score. It is not clear what the authors did with the rest of the students. Students' ages ranged between 21 to 45 years old, but it was not made clear whether the two groups were also balanced in terms of age, as age could be an important

factor in students' growth in those skills. The researchers do not also state whether the two groups were taught by the same instructor or different instructors. If classes are taught by the same instructor, there might be unintentional diffusion of treatment and if classes are taught by different instructors, then teacher effect cannot be controlled for. For this reason, the number of clusters need to be large so that any differences in teacher effect can be cancelled out. For this reason, the study was rated 1*.

In summary, there is indicative evidence that explicit teaching of general critical thinking skills can improve English language learners' critical thinking skills. Although the evidence is not strong due to the small sample in most of the studies and attrition in some, it is the most promising approach with the most number of 2* studies showing positive effects. The prevalence of so many poor quality studies in this field, with many having no proper comparison groups, or randomising two intact classes and high attrition suggests an urgent need for large-scale well-designed randomized controlled trials where attrition is minimized.

The rest of the studies (Moore 1995; Turuk 2011; Chason et al. 2017) were rated 0. Moore and Chason et al. were single-group design and thus have no counterfactuals. In Moore's study students were Malaysian students studying in America. With no comparison group it is difficult to tell whether the gains in critical thinking is the result of the intervention, natural maturation or simply the experience of being immersed in a different culture. In Turuk's study only 16 of the original 27were analysed. A number of students dropped out after the pre-test and during the intervention. Apparently some students dropped out because they found the course materials challenging. Therefore, including those who remained is likely to get a skewed result since those who are not likely to do well have excluded themselves from the analysis. The weak design (having no comparison group) plus the very small sample size and high attrition – all meant that there is low credibility in their findings, hence the low rating.

5.2 Approaches with little of evidence of effectiveness

Besides general critical thinking skills instruction, all the other approaches showed little or no evidence of effectiveness despite almost all claiming positive results. These include strategies such as debate, use of self/peer assessment and feedback, use of literary and narrative texts, brainstorming techniques, scaffolding and other active learning strategies (e.g. collaborative writing, journal writing, and dialogic thinking).

5.2.1 Debate

The evidence for debate as a teaching strategy to develop students' critical thinking is weak. Only three studies evaluated the use of debate as an instructional strategy for English language learners in higher education. Of the three, two showed positive effects, and the best study was rated 1.5* (Tous, Tahriri & Haghighi, 2015). A third study by two of the same authors (Tous & Haghighi, 2016) showed no effects but it compared the results for males and females and so was not relevant to the review question. All were rated low in strength of evidence.

Table 3: Quality and impact detail: studies focused on debate (N = 3)

Author(s) + Year	Smallest cell	Attrition	Effect size	NNTD	Quality
Tous, Tahriri and Haghighi (2015)	44	Not reported	1.01 (calculated by reviewers)	44	1.5*
Yang and Gamble (2013)	31	Not reported	0.74 (calculated by study authors)	23	1*
Tous & Haghighi (2016)	Not applicable	Not reported	Not applicable*	-	0

* single-group design – no comparison of gainscores

Tous, Tahriri and Haghighi (2015) examined the effect of debate training on the reading comprehension of 88 students. This was a quasi-experiment where 88 participants were selected by convenience sampling and "grouped" (authors' term) into control and experimental groups. It is not clear if allocation was randomized but since it was not described as such we assume that it was not. Experimental group was trained using the Meeting-House Debate Strategy where they were taught skills in presenting arguments and challenging flaws in the opponents' arguments. The control group received the usual instruction based on the traditional lecturing technique. Teaching was in English. Critical thinking skills were assessed before and after the intervention using the Read Theory Critical Reading Comprehension Test (RTCRCT) and the Persian version of the CCTST test. The study reported strong positive effects on both the CCTST and the RTCRCT tests, but this was an analysis of correlation rather than a comparison of gain scores. The analysis was not clearly explained and it was also unclear how groups were assigned. There was no report of attrition or missing data. Also the study spanned

over only one month, so there is the threat of students becoming familiar with the test, and it is questionable whether the short duration would result in such a big gain. Due to the ambiguity in reporting and the short duration of the intervention, the study was rated 1.5*.

Yang and Gamble (2013) reported a huge effect of integrating debate in the EFL curriculum on students' level of critical thinking. This was a cluster randomized trial of only two intact classes consisting of 68 students. Since there were only two classes, allocating one class at random to treatment condition cannot be considered technically as randomization. Effectively the sample size is only two clusters. Clusters usually have inherent qualities so students in each cluster might be similar to each other but different from the students in the other cluster. The two groups were taught by the researcher in the study. There is therefore a possibility of teacher expectation, which could bias the results in favour of the experimental group. It is also not mentioned whether the two raters who graded the essays were blinded. If the raters knew which group the students belonged to there is a likelihood of bias**.** There was also no report of missing data or attrition. Given the short duration of the intervention (8 weeks), there is a possibility that students may become familiar with the test. This is especially so if the treatment students have been exposed to similar elements in the intervention as those in the test.

The evidence for debate as an approach to foster critical thinking of English language learners is not strong largely because of the small number of studies (so lack of replication), the very small sample and the inadequate reporting of key information.

5.2.2 Assessment techniques as an instructional approach

A total of five studies evaluated the use of assessment techniques on students' critical thinking skills, and all five reported a positive outcome. Assessment techniques include a variety of strategies like conferencing, peer-review, peer-evaluation, and self-evaluation.

Three used standardized tests of critical thinking. Two of the studies used the Watson-Glaser Critical Thinking Appraisal (WGTA) as a pre-test and a post-test, one used the California Critical Thinking Skills Test, another used the Cornell Critical Thinking Test, and one also used an argumentative essay.

All the studies were rated poor due to major flaws in design, such as using intact groups and no control for confounding variables. Thus there is very little evidence that this strategy as used

in the studies in this review works in improving critical thinking. None of the studies involved random allocation of participants into treatment conditions, and none reported attrition. All were very small scale (see Table 4). One was given a zero rating (Iraji et al., 2016) because the reporting was found to be inadequate. The outcome was performance in an argumentative essay. It is not clear how the rating was done and whether the raters were blinded to avoid bias since knowledge of treatment conditions can unconsciously affect one's judgement. The sample size was small (N = 36) and there was no report of duration of the intervention nor the number of essays students had to write.

Table 4: Quality and impact detail: studies focused on assessment techniques (N = 5)

Author(s) + Year	Smallest cell	Attrition	Effect size	NNTD	Quality
Daud, Gilmore, and Mayo (2013)	24	Not reported	Not enough data provided	-	1*
Jafari and Yavari (2014)	30	Not reported	0.72 (calculated by reviewers)	22	1*
Jafari, Yavari, and Ahmadi (2015)	25	Not reported	0.59 (calculated by reviewers)	15	1*
Kahrizi, Farahian, and Rajabi (2014)	20	Not reported	0.34 (calculated by reviewers)	7	1*
Iraji, Enayat, & Momeni, 2016	18	Not reported	1.88 (calculated by reviewers)	34	0

Four studies reported positive effects were rated 1*. **Daud, Gilmore, and Mayo (2013)** examined the use of peer review, self-evaluation and peer evaluation on the development of students' critical thinking skills and writing ability. Students forming 4 intact groups (n = 99) enrolled in an English for Academic Writing course participated in the study with one group serving as control and three as experimental with one focusing on peer review, one on self-evaluation, and one on peer evaluation. With only 99 students divided into 4 groups, there could only be about 20 in each. Since the students were not randomly allocated, inherent differences between groups can still exist. For example they may differ by age or prior

attainment. Impact was measured by correlating the CCTT-X post-test scores with their final term paper scores. The researchers did not provide information on how the final term papers were graded and whether raters were blinded. Not blinding raters could bias the results. It is not clear why a simple analysis comparing the gains from pre- and post-tests was not employed. The authors reported significant correlations between critical thinking skills and academic writing ability for the peer review and peer evaluation groups, suggesting that these two assessment techniques were more effective than self-evaluation and self-review. However, comparing scores on the critical thinking test with the performance on the term papers does not provide a credible measure of effectiveness since students who score highly on critical thinking are likely to also write well. Data analyses were presented with no standard deviation, making it impossible to calculate the effect size. The study received a 1* rating.

Jafari and Yavari (2014) examined the effect of conferencing on students' critical thinking, using a pre-test and post-test design on two groups of learners (n = 60). A lapse of only seven weeks between the pre-test and the post-test might have resulted in students becoming familiar with the test, which might have biased the results in favour of the treatment group as they have just been exposed to the rubrics of critical thinking in the pre-test, which closely aligns with the intervention. The participants were in two classes and one class was "selected" to receive the intervention. Participants were clearly not individually randomised. This means that the number of cases would effectively be two. Although a pre-test was taken to establish equivalence, unobservable differences may still exist between the classes, for example, in terms of teacher quality. The paper was very sparse in information. We do not know if the two classes were taught by different teachers or not. The authors claimed that because "None of the candidates knew that they were part of a research project", it was a "kind of randomization" (p. 154). The outcomes were measured using the Persian version of the WGCTA although instruction was in English. There was also little information about the intervention. All we know is that treatment students were given time to speak about their problems and then they were given feedback by their teacher in the conferences. It is not clear what kind of feedback was given to students in the conference sessions or the number of sessions delivered. The authors mentioned that while the experimental group got feedback, the control students had to write essays but were not given any kind of oral or written feedback from the teacher or their peers. This is equivalent to withdrawal of teaching for the control the students. It is often the case that if you teach someone more of something they know more about that thing. The control students are therefore disadvantaged as there is no support for learning for them. The study

reported a huge impact but this could be attributed to the small sample size (n = 60). Results for each of the subsections of the post-test were presented, but for the pre-test only a composite score was given.

A later report by two of the authors in the previous study (**Jafari, Yavari, & Ahmadi, 2015)**, suggested that self-assessment had a positive effect (ES = +0.59) on the critical thinking and language proficiency of students. The study involved 50 students from two intact classes. One class practiced self-assessment while the other class served as the control. As the participants were not randomly assigned to treatment conditions, the groups could be different from the outset. For example, one class could be taught by a more effective teacher (not clear if the two classes were taught by the same teacher or not), or could be different in terms of prior attainment. As before, the authors argued that because the candidates did not know that they were part of a research project, this meant that they were in random groups (p.146). There was little information about what the intervention was and what the control students did. It is also not clear whether students in the experimental group assessed themselves orally or in written form, and whether they assessed the essay structure, logic, or language. It is possible that in these two studies (Jafari & Yavari, 2014; Jafari, Yavari, & Ahmadi, 2015), teachers may be teaching to the test. If the control group was not given any support for learning and left to their own devices, this is tantamount to withdrawal of instruction. Therefore any comparisons between the two groups would be unfair. The poor reporting, small sample and lack of random allocation to treatment conditions all meant that the findings of the study are not reliable.

Another study which evaluated the impact of self-assessment (**Kahrizi, Farahian, & Rajabi, 2014)** also reported a big effect. Participants were 40 students from three classes selected based on a TOEFL test. The self-assessment group was given a checklist focusing on organization, content, vocabulary, language use, and mechanics. In addition to the small number of cases, the process of randomization was not explained clearly, and attrition was not reported. It is not clear whether individuals or groups were randomized.

In summary, there is no evidence that assessment techniques as an approach to enhance students' critical thinking is effective despite huge effect sizes cited. All the studies that evaluated this approach were small in scale and did not involve randomizing individuals. Randomly picking one class to receive an intervention is not proper randomization, and

comparing groups who receive instruction on critical thinking while withdrawing instruction and support for learning for the control group cannot be seen as a fair comparison.

5.2.3 Literary and narrative texts

Another instructional approach is the use of literary and narrative texts to enhance students' critical thinking skills. There is no evidence that this approach is effective. Six studies examined this approach with four receiving a rating of 0 due to their weak designs and poor reporting and two receiving a rating of 1* (see Table 5). All reported positive outcomes. We discuss only the two studies that were rated 1*. The other three were so poor that they would not contribute to the evidence.

Table 5: Quality and impact detail: studies focused on literary and narrative texts (N = 6)

Author(s) + Year	Smallest cell	Attrition	Effect size	NNTD	Quality
Fatemi (n.d.)	47	Not reported	0.99 (calculated by reviewers)	47	1*
Khatib and Alizadeh (2012)	17	Not reported	Not enough data provided	-	1*
Arslan & Yildiz (2012)	Not applicable	Not reported	Not applicable*	-	0
Khatib & Janpour (2012)	15	Not reported	0.99 (calculated by reviewers)	15	0
Pashangzadeh, Ahmadian, & Yazdani (2016)	27	Not reported	0.89 (calculated by reviewers)	24	0
Khamkhong (2018)	Not applicable	0%	Not applicable*	-	0

* single-group design – no comparison of gainscores

Fatemi (n.d.) examined the impact of literary narratives using a quasi-experiment. A total of 105 EFL (English as a Foreign Language) university students from two different universities were selected for the trial. Students from one university taught by the researcher formed the

experimental group, while those in another university formed the control. Outcomes were measured using the Persian version of WGCTA. Students in the experimental group were asked questions that encouraged the use of critical thinking skills while reading narrative texts in class, and the control group had essays to read. Although the author states that the two groups are balanced in language proficiency, background, age, and critical thinking, the two groups are from two different universities, so they could be different in other unobservable characteristics. The experimental group was taught by the researcher but nothing was mentioned about the teacher who taught the control group. It is possible that the researcher could be teaching to the test (especially if they knew the contents of the test). The huge effect size cited (ES = +0.99) could be due to the small sample size or, more likely the result of teaching to the test. There was also no report of attrition or missing values.

Another study that examined the effect of using literary texts (**Khatib & Alizadeh, 2012)** was a two group pre-post design using the WGCTA as the test instrument. Thirty-four students (out of 46) were selected based on the results of the pre-test and divided into two groups. Both groups were taught critical thinking, but the experimental group used literary texts while the control group used non-literary texts usually found in academic textbooks. Although the author claimed that the participants were "randomly assigned" to two groups, it is not clear how this was carried out as they also stated that they wanted to have equal numbers of male and females in each group. Was it stratified or was it proportional randomization, or was it ad hoc? It appears that many researchers confused ad hoc allocation with randomization. It is also not clear if the two groups were taught by the same instructor. The analyses were so badly reported that it was hard to make out what the effect size would be. Instead significant tests (t-tests) were used, which are inappropriate.

Given the very weak studies so far, there is little evidence that literary and narrative texts are an effective way to enhance critical thinking skills of English language learners.

5.2.4 Brainstorming techniques

Another strategy that has been tested is brainstorming techniques. This includes a strategy called concept mapping. Brainstorming is a technique to help students generate ideas and relate ideas to each other. The two studies that evaluated this approach both reported a positive outcome. Both used the WGTCA, but one compared two groups of students (which could be

different at the outset), and the second study privileged the treatment group by giving them additional support. Both were given a rating of 1*.

Table 6: Quality and impact detail: studies focused on brainstorming techniques (N = 2)

Author(s) + Year	Smallest cell	Attrition	Effect size	NNTD	Quality
Ghabanchi and Behrooznia (2014)	25	Not reported	0.75 (calculated by reviewers)	19	1*
Khodadady and Ghanizadeh (2011)	18	Not reported	1.21 (calculated by reviewers)	22	1*

In **Ghabanchi and Behrooznia's (2014)** study, 54 university students from two intact groups on a reading course were involved in the trial. This was a two-group, pre- post quasi-experimental study. Participants were not randomised to treatment conditions but conveniently assigned. Therefore the number of cases is not 54 but 2 clusters. The two groups were taught by the same teacher, who was also the researcher, using the same material with the only exception that brainstorming was practised in the treatment group. There is therefore a threat of selection bias as the two clusters might be completely different from each other, and students forming each cluster might share similar qualities. As with most other studies there was no report of attrition or missing values. The study reported a huge effect size (ES = +0.75). The analyses were badly presented. For example, the mean pre- and post-test scores for the two groups were not presented in the tables. Instead the results of significant tests were used to show that the two groups were different. This was despite having no random samples.

Another study looked at the effect of concept mapping (a brainstorming technique) on 36 EFL students' critical thinking (**Khodadady & Ghanizadeh, 2011)**. The TOEFL test was administered to all students to ensure that they had the same proficiency level. The groups were assigned to treatment conditions based on their pre-test. In other words, allocation was not random even though the authors claimed that the students were randomly assigned to the two groups. The intervention was delivered in 22 two-hour sessions. In each session, students in the experimental group were given a reading passage and were asked to construct a concept map at home using the software C-map tools. The maps were then discussed in class the

following day. The control group was not assigned any homework. The same instructor taught the two groups. As the experimental students were required to do the concept maps at home, other variables could have affected the study. For example, students could have been given extra help from parents, siblings, or friends or could have done additional reading up. It is therefore not possible to rule out the influence of other extraneous factors. This could be controlled if the activities were completed in class. There is also a possibility of a Hawthorne effect as the use of the software for generating concept maps is a novel idea. Attrition rate was not reported.

Although brainstorming as a technique to teach logic and critical thinking might be a useful strategy to help students generate ideas, the evidence of its effectiveness is weak. There were only two studies that evaluated this approach. Both were small scale, and both involved unclear randomisation.

5.2.5 Journal writing

Journal writing is another approach used to develop critical thinking skills of English language learners. Two studies were identified using this approach and both reported a positive outcome. One was rated 0 and the other given a 1* (Table 7).

Table 7: Quality and impact detail: studies focused on journal writing (N = 2)

Author(s) + Year	Smallest cell	Attrition	Effect size	NNTD	Quality
Khatib, Marefat, and Ahmadi (2012)	9	Not reported	Not enough data provided	-	1*
Shaarawy (2014)	7	Not reported	Not enough data provided	-	0

Shaarawy's (2014) study was a quasi-experiment involving 56 first year university students (33 in the experimental group and 23 in the control group). This was rated zero because of inadequate reporting, very small sample size (N =23). It was not clear how the groups were formed, but it is very likely that they were in two intact classes on the same course. Both groups were taught the same syllabus by the same teacher who was also the researcher. The only difference was that the intervention group was given an additional weekly journal writing

exercise where writing prompts were given based on Bloom's taxonomy of cognitive skills. Critical thinking was measured using a researcher-developed tool based on Bloom's taxonomy of cognitive skills. As the test is related to the intervention, which was not given to the control group, it cannot be considered a fair test. Also, as in most other studies in this review, the impact of the intervention was calculated using t-tests despite no randomization. Final analyses were conducted on only 16 experimental students who had completed all the seven journal writing exercises and who had pre- and post-tests scores. Only seven of the control students with pre- and post-test scores were included in the final analyses. This represents an attrition of 55%. Students who completed all the writing exercises may be different in terms of motivation and prior skills compared to those who did not. This is thus a bias in the selection of intervention students.

Khatib, Marefat, and Ahmadi (2012) examined the effect of keeping audiotaped and written dialogue journals on students' critical thinking. Students from three intact classes were included in the study (two experimental classes and one control class). The two experimental groups were instructed to keep journals, with one group keeping written journals (n = 19) while the other group kept audiotaped journals of 5 to 10 minutes (n = 9). Students were encouraged to reflect on any topic of their choice in their journal on a weekly basis over 19 sessions. The instructor provided feedback on their journal entries. The control group (n = 12) had regular class activities with no special tasks. All three groups were taught by the same instructor introducing the possibility of diffusion. Critical thinking was assessed using the Persian version of the WGCTA although instruction was given in English. The authors concluded that students using journal keeping (both written and audiotaped) performed better than the control and there was no difference between written and audiotaped journal keeping in terms of effectiveness. This study was rated 1* because of the very small sample (under 50), unclear reporting of attrition rate and the misuse of significant testing in comparing effects. We do not know how many students were there at the beginning. We only know that all the 33 students who completed the WGCTA were included in the final analysis. As before no standardized effect size was calculated. ANOVA and t-tests based on significant testing were used to compare the results of the three groups even though the samples were clearly not randomized. The authors explained that the students were placed in the three classes based on their oral and written placement tests, suggesting that the three groups were already different at the outset.

5.2.6 Scaffolding

Scaffolding as a strategy to enhance critical thinking skills has been evaluated by only two studies (Table 8). One was given a rating of 1* (**Sokol et al., 2008**) and reported a positive outcome and one was given a rating of 0 (**Hurte, 2004**) and reported no effect because it was not a test of the effectiveness of the scaffolding strategy, but a comparison of scaffolding with the Cognitive Enrichment Advantage (CEA) approach. Both groups registered a decline between pre- and post-test, with the scaffolding group showing a bigger decrease. This suggests that the scaffolding strategy is less effective than the CEA approach. Participants were first year university students who were matched in pairs and randomly assigned to treatment conditions. Given that there were only 36 students, the matched pair assignment meant that the number of cases was effectively only 18. Moreover both groups received two weeks of direct instruction in critical thinking. The absence of a control group, the lack of individual randomization and the fact that the instructor was also the researcher all weaken the evidence. Hence it was given a zero rating.

Table 8: Quality and impact detail: studies focused on scaffolding (N = 2)

Author(s) + Year	Smallest cell	Attrition	Effect size	NNTD	Quality
Sokol, Oget, Sonntag, and Khomenko (2008)	27	Not reported	Not enough data provided	-	1*
Hurte (2004)	18	Not reported	Not applicable*	-	0

* Not a test of critical thinking skills strategy but a comparison of two approaches

The other study by Sokol et al. (2008) was a quasi-experiment comparing 54 students from one school (4 classes) with 27 students from another school (2 classes). The intervention, known as the Thinking Approach integrates inventive thinking skills instruction in foreign language teaching. The teacher's role was to scaffold learners who had to build models by responding to certain specific tasks. The experimental students had 5 hours of English per week while the control group received only 3 hours per week. The two groups were from two different schools, one in the capital city and one in a town, which might have also resulted in biased results. As

the groups were not randomly allocated to conditions there may be systematic differences between them. It is therefore not possible to rule out other confounding effects. The authors acknowledged that the groups also differed in terms of proficiency level and teacher expertise. All these pose threats to the internal validity of the study. Outcomes were measured using an inventive thinking test which is closely aligned with the intervention. Moreover the test was graded by only one rater who was not blinded. Attrition was not reported nor was the effect size. Instead a comparison of groups using t-test was conducted. This is an inappropriate analysis as the sample is not random. Significant tests cannot be used for non-random samples. All these rendered the results untenable.

Therefore, we can conclude that there is no evidence that scaffolding is effective in developing critical thinking skills in English language learners.

5.2.7 Active learning strategies

Other active learning strategies identified in this review include the use of collaborative writing and dialogic thinking. Only three studies were found that examined those strategies, and all reported a positive outcome (Table 9).

Table 9: Quality and impact detail: studies focused on active learning strategies (N = 3)

Author(s) + Year	Smallest cell	Attrition	Effect size	NNTD	Quality
Kusumoto (2018)	62	17.7% (29)	0.03 (calculated by reviewers)	1.86	2*
Rashtchi (2007)	36	Not reported	Not enough data provided	-	1*
Fahim & Mirzaii (2013)	21	Not reported	1.24 (calculated by reviewers)	26	0

In a quasi-experiment involving 162 participants, **Kusumoto (2018)** examined the use of active learning on students' level of critical thinking over a period of two semesters. Two classes taught by two different teachers were compared. This reduces the credibility of its findings since the two classes may be inherently different and any difference between groups cannot be attribute to the intervention. Some students were also enrolled in English courses in

the same year, which might have contributed to students' growth in critical thinking. The researcher excluded 29 students with missing scores from the analysis. Students who comply till the end might be different from students who miss the test, so the researcher should have presented the pre-test scores of those missing students to make sure that they are not any different from those who complied. The number needed to disturb the finding is 1.86 which would be rounded to 2. This means that 2 counterfactual cases would be needed in order to change the findings. The number is low compared with 29 missing cases. The effect size as calculated by the authors of this review is 0.03 which is considered a very small effect size, indicating that there is no difference between groups. Therefore, the study does not provide strong evidence that active learning strategies could enhance students' level of critical thinking. The study was given a rating of 2*.

Fahim and Mirzaii (2013) evaluated the use of dialogic thinking where the experimental students received dialogic critical thinking training in addition to argumentative writing instruction. Control students were trained only in argumentative writing. Participants were 43 male EFL learners (out of 48) from four classes who scored ±1 standard deviation of the mean score in an argumentative essay. Two classes were randomly assigned to experimental condition (n = 21) and two to control (n = 22). Post-test analysis included only 42 students. It is not clear what happened to the 43rd student. The study showed a huge gain between pre- and post-test on a researcher-developed English written test (ES = 1.45, calculated by the reviewers). It is unclear whether the researchers were also the teachers teaching the experimental classes and whether they marked the tests as well. If so, then there could be a teacher expectation effect. This study was rated 0 due to the poor reporting, small sample size (n = 4 clusters), the use of a researcher-developed test, and lack of blinding of markers. This again proves the point that Bob Slavin made in his blog (Slavin, 2018) about the 10-foot man.
Rashtchi (2007) examined the effect of collaborative writing. Participants were 74 students from an Islamic university in Tehran who scored ±1 standard deviation of the mean score in the Comprehensive English Language Test (CELT), with 38 in the experimental and 36 in the control group. Interestingly these students were pre-randomised before the test from a total of 90. This meant that sixteen students were excluded after randomisation, representing an attrition of 18% even before the trial started. Experimental students received 14 sessions of cooperative writing while students in the control group wrote individually with the instructor giving feedback to both groups at the end of each session. The researcher was the instructor of both groups. This means that there is a possibility of bias even if unintended. Critical thinking

was assessed using WGCTA. The very small sample size (n = 4 clusters), the very high attrition after randomisation, the very poor reporting, misuse of significant tests and the fact that the researcher was also the instructor meant that the evidence is untenable.

6. Limitations

As with all reviews it is possible that some studies may have been missed. For example, the parameters set for the search included only articles published in English, from 1990 to 2018. This may have excluded relevant materials that are outside these parameters. The key issue is whether including those studies would have altered the findings. This review searched specifically for studies about teaching critical thinking to English language learners in higher education. Therefore studies about effective approaches to teaching critical thinking skills for English native speakers were not included. We acknowledge that these could shed light on some of the more effective approaches. This could be explored in a future review.

7. Conclusions

Several strategies for developing critical thinking skills have been tested, and almost all claimed positive effects. No studies reported negative effects of teaching critical thinking. Therefore, we could not identify any approaches that were not effective. It is possible that this could be due to publication bias where positive results are more likely to be published or where researchers are more likely to publish if they found positive effects. It may also be the case where researchers are so keen to find positive results that they report only the positive results.

Almost all the studies in this review are very small-scale and have serious methodological flaws. Of the 36 studies that were synthesized, thirteen of them were given a 0 rating. Seventeen were given a rating of 1*. The best studies in this review were rated 2* (n = 5), and 1.5* (n = 1). No studies were rated above 2*. Therefore there is little evidence that any of the approaches actually works.

However, the approach involving instruction in general critical thinking skills looks the most promising, but more large-scale and robust evidence is needed to confirm its effect. This approach has been evaluated by the biggest number of studies with the highest number of studies rated 2* (the best rating in this review). Overall the evidence is weak due to the quality of the studies.

7.1 Common problems identified in this review

No study in this review was rated above 2*, suggesting that research in this area is still rather premature. Of the 1,794studies, 1,000 were found via handsearching google scholar mostly in journals that are not international in scope and are invariably of poor quality. Almost all were very small scale, conducted by researchers who were themselves the instructors using students in their own institution or classes. Most of the approaches were evaluated by fewer than three studies. The small-scale and the lack of replication meant that it is not possible to say for sure which approach is really effective.

Also a large number of studies involved ad hoc randomisation or pseudo-randomisation where two classes were "randomly" picked to receive the intervention. It is also the case that in a large number of studies the experimental group was given additional support (in addition to the regular lessons), while control students were not, and in some cases instruction was even withdrawn from the control group. Comparing students who were given extra help with those who had no help at all is not a fair comparison.

A large number of studies used standardised tests that were translated into the native language of the students even though the intervention was delivered in English. Critical thinking requires the ability to make arguments, understand logical fallacies, question assumptions, make warranted conclusions and offer alternative explanations. How closely these skills can be translated in another language is questionable. Some common words like evidence, reliable/unreliable, take for granted, prediction, unstated assumption in the Cornell CTT and WGCTA test might be an obstacle if students do not know their equivalence in their own native language. It makes sense that if the study was conducted in an ESL/EFL context and the intervention was delivered in English, then the test instrument should be English. The argument often put forward for using the translated version is that standardized tests are culturally biased. But translating the test into another language may remove some of the subtle nuances which are particularly relevant in critical thinking.

Many studies in this review have reported the short duration of intervention as a main barrier to students' growth in critical thinking. This suggests that a longer period may be needed for effects to be realised as critical thinking skills require time to develop.

Another issue faced in this review is the absence of a single agreed-upon definition for CT, which makes comparison of studies difficult as different studies may be measuring different things. Although the majority of studies used standardised tests like the Watson Glaser Critical Thinking Appraisal (WGCTA), California Critical Thinking Test (California CCT) and the Cornell Critical Thinking Test (Cornell CTT), a few other studies used bespoke or adapted versions of the test or researcher-developed writing tests.

Another prevalent practice is the misuse or misinterpretation of significant tests. Significant tests are not appropriate for quasi-experimental studies using convenient samples, or matched groups with no random samples. Even when there is proper randomisation, any missing data or attrition would have rendered the sample non-random as missing cases are rarely random. In some studies, students who did not complete the post-test were excluded from the analysis. Significant tests are based on the premise that there is complete randomisation. And even if there is complete randomisation significant tests are still not appropriate because null hypothesis significant testing (NHST) states that assuming there is no difference between groups how likely are we to obtain data as extreme as observed. The answer that most researchers want is: given the data how likely is there a difference between groups. Unfortunately, significant tests do not and cannot answer this question. All this shows that there is much still to be done in research in this area.

7.2 Recommendations for implementation of critical thinking strategies

Longer exposure to critical thinking instruction

The short duration of the intervention is cited in a number of studies as a barrier to successful implementation. Most of the studies in this review involved teaching critical thinking over a semester (between 12 to 16 weeks). Critical thinking comprises a set of complex skills, which are often not familiar to EFL/ESL learners. Constant reinforcement and application of those skills is needed to develop those skills. Therefore, we suggest that evaluations of critical thinking skills approaches should be conducted over at least one semester for effects (if any) to be realised.

Training of teachers

To teach critical thinking the teachers themselves must be able to think critically. None of the studies reviewed discussed teacher preparation or described how it took place. This is perhaps because, in most cases, the researchers are themselves the teacher. In practice, teachers

delivering such interventions must be adequately trained. In some studies, it was not even clear whether the researcher or the teacher taught the different groups and whether the groups were taught by one or more teachers. Where more than one teacher was involved, there was no report about how or whether teachers were trained. No process evaluation was carried out to ensure that the intervention was delivered as intended. None of the studies explained how consistency of delivery across groups was maintained. Our second recommendation, therefore, is intensive training of teachers to ensure that teachers have the required thinking skills themselves and the competence to deliver the instruction.

7.3 How can research in this area be improved?

Given the large number of small-scale studies, often carried out by researchers themselves involving their own students, what is now needed for clearer evidence is well-designed, large-scale, independently evaluated randomised controlled trials using standardised tests of CT in the language of instruction. Our recommendations are:

- More rigorous and robust evaluations of the impact of critical thinking approaches. Ideally they should be large-scale (over 100 in each intervention arm) and conducted by independent evaluators.
- Participants should be properly randomised, preferably individually. Where classes or schools are randomised, there should be a big enough number to ensure that the groups are equivalent.
- There should be replications of the better positive studies. For example, the general critical thinking skills approach and debates could be tested in an efficacy trial.
- Assessments should be by independent assessors who are blind to treatment allocation.
- The licensed version of the test instrument in the language of instruction should be used to avoid problem of language transference. This also minimises the possibility of researchers teaching to the test if an adapted or modified version is used.
- Where approaches involve the use of unconventional strategies such as computer software or video recording (as in the concept map approach), an alternative innovative treatment should be used to ensure that any impact is not due to the novelty effect.
- All use of signi,ficant test and its variance should be banned. They are misleading at best and harmful at worst. They lead to invalid and therefore potentially damaging research outcomes (Cohen, 1994; Trafimow & Rice, 2009; Colquoun, 2014, 2016;

Perezgonzalez, 2015; Gorard, 2016). P-value in significant tests does not tell us whether there is a real difference between the groups compared. This is a common misinterpretation of significant test (Kline, 2004). The irony is that teachers/researchers of critical thinking themselves fall for the common fallacies of significance tests. Instead calculation of effect size should be used. This is the difference in the mean gain scores between the comparison groups. Data analyses should include basic information like the mean pre-test scores and the mean post-test scores of the two groups being compared as well as the standard deviation. It is good practice to also report any missing data, missing values and attrition.

- Where there is missing data, attrition or non-compliance, both intention-to-treat and compliance average causal effect analysis should be used.
- Process evaluations should form part of the evaluation especially in complex interventions so that if the programme works we can identify the mechanism that brings about change, or factors that are necessary for successful implementation. And if the programme fails, process evaluation is useful in identifying those factors that may have hindered effective implementation.
- Clear, complete and transparent reporting is necessary if research in this field is to advance.

In general, we believe that the ability to think critically is a very useful skill and should be taught. The review has shown some indicative evidence that explicit teaching of critical thinking is possible and may be effective for some approaches.

References

Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically a meta-analysis. *Review of Educational Research, 85*(2), 275–314. DOI: 10.3102/0034654314551063

Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R., & Zhang, D. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of Educational Research, 78*(4), 1102-1134.

Akbari, M., Seifoori, Z., & Ahour, T. (2017). Enhancing comprehension and production of argumentation through critical thinking awareness-raising. *Linguæ &, 16*(2). doi: https:doi.org/10.7358/ling-2017-002-seif

Andrews, R. (2015). Critical thinking and/or argumentation in higher education. In M. Davies & R. Barnett (Ed.), *The Palgrave handbook of critical thinking in higher education* (pp. 49-62). London: Palgrave Macmillan.

Arslan, R. Ş., & Yildiz, N. (2012). Enhancing critical thinking at the tertiary level through a literature-based critical thinking program. *Journal of the Cukurova University Institute of Social Sciences, 21*(2), 19-36.
Available at https://www.researchgate.net/publication/258028491_ENHANCING_CRITICAL_THINKING_AT_THE_TERTIARY_LEVEL_THROUGH_A_LITERATURE-BASED_CRITICAL_THINKING_PROGRAM

Arum, R., & J. Roksa J. (2011). *Academically adrift: Limited learning on college campuses*. Chicago: University of Chicago Press.

Association of American Colleges and Universities. (2004). *Liberal education outcomes: A preliminary report on student achievement in college.* Washington, DC.

Association of American Colleges and Universities. (2015). *Falling short? College learning and career success*. Washington, D.C. Retrieved from http://www.aacu.org/sites/default/files/files/LEAP/2015employerstudentsurvey.pdf

Atkinson, D. (1997). A critical approach to critical thinking in TESOL". *TESOL Quarterly*, 31(1),71–94.

Bailin, S., Case, R, Coombs, J. R., & Daniels, L. B. (1999). Common misconceptions of critical thinking. *Journal of Curriculum Studies, 31*(3), 269-283.

Behar-Horenstein, L. S., & Niu, L. (2011). Teaching critical thinking skills in higher education: A review of the literature. *Journal of College Teaching & Learning (TLC),*

8(2), 25-42.

Blackmore, J. (2001). Universities in crisis? Knowledge economies, emancipatory pedagogies, and the critical intellectual. *Educational Theory, 51*(3), 353-370.

Chason, L., Loyet, D., Sorenson, L. & Stoops, A. (2017). An approach for embedding critical thinking in second language paragraph writing. *TESOL Journal, 8*, 582-612. doi:10.1002/tesj.288

Coe, R. (2002). It's the effect size, stupid: What effect size is and why it is important. Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, England, 12–14 September. Retrieved from www.leeds.ac.uk/educol/documents/00002182.htm

Cohen, J. (1994). "The earth is round (p<.05)". *American Psychologist, 49*(12), 997-1003.

Coil, D., Wenderoth, M. P, Cunningham, M., & Dirks, C. (2010). Teaching the process of science: Faculty perceptions and an effective methodology. *CBE Life Sciences Education, 9*(4),524-535.

Colquoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society Open Science, http://rsos.royalsocietypublishing.org/content/1/3/140216.

Colquoun, D. (2016). The problem with p-values. Aeon, https://aeon.co/essays/it-s-time-for-science-to-abandon-the-term-statistically-significant

Cook, T. D. (2002). Randomized experiments in educational policy research: A critical examination of the reasons the educational evaluation community has offered for not doing them. *Educational Evaluation and Policy Analysis*, *24*(3), 175-199.

Cook, T. D., & Shadish, W. R. (1994). Social experiments: Some developments over the past fifteen years. *Annual Review of Psychology, 45*(1), 545-580.

Daud, N. S. M., Gilmore, A., & Mayo, H. E. (2013). Exploring the potency of peer evaluation to develop critical thinking for tertiary academic writing. *World Applied Sciences Journal, 21*, 109-116. Available at https://www.idosi.org/wasj/wasj21(SLTL)13/14.pdf

Davidson, B. W., & Dunham, R. L. (1997). Assessing EFL student progress in critical thinking with the Ennis-Weir Critical Thinking. *JALT Journal*, 19(1), 43-57. Available at https://files.eric.ed.gov/fulltext/ED403302.pdf

Davies, M. (2003). *A Cautionary Note about the Teaching of Critical Reasoning.* Paper presented at the 'Learning for an Unknown Future' Conference, Higher Education

Research and Development Society of Australasia (HERDSA), Christchurch, New Zealand.

Davies, M. (2011). Introduction to the special issue on critical thinking in higher education. *Higher Education Research & Development, 30*(3), 255-260, DOI:10.1080/07294360.2011.562145

Dewey, J. (1910). *How we Think*. 1991 ed. Buffalo, NY: Prometheus Books.

Dong, Y. (2017). Teaching and assessing critical thinking in second language writing: An infusion approach. *Chinese Journal of Applied Linguistics*, *40*(4), 431-451. DOI:10.1515/cjal-2017-0025

Driver R., Newton, P., & Osbourne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education*, *84*(3), 287-312.

Eurydice. (2011). *Science Education in Europe: National Policies, Practices and Research*. Brussels: EACEA.

Facione, P. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction.* Millibrae, CA: The California Academic Press.

Fahim, M., & Mirzaii, M. (2013). Improving EFL argumentative writing: A dialogic critical thinking approach. *International Journal of Research Studies in Language Learning*, *3*(1), 3-20. Available at http://consortiacademia.org/wp-content/uploads/IJRSLL/IJRSLL_v3i1/313-1727-1-PB.pdf

Fatemi, A. H. (n.d.). Incborporating critical thinking into EFL curriculum through fictional -narrative based reading and awareness of consequences technique. International Conference "ICT for Language Learning". Available at https://conference.pixel-online.net/conferences/ICT4LL2012/common/download/Paper_pdf/58-QIL08-FP-Hosseini-Fatemi-ICT2012.pdf

Ghabanchi, Z., & Behrooznia, S. (2014). The impact of brainstorming on reading comprehension and critical thinking ability of EFL learners. *Procedia - Social and Behavioral Sciences*, *98*, 513-521.

Gimenez, M. E. (1989). Silence in the classroom: Some thoughts about teaching in the 1980s. *Teaching Sociology***,** *17*, 184-191.

Gomez, J.C. (2010). The impact of structured reading lessons on the development of critical thinking skills. *Electronic Journal of Foreign Language Teaching, 7*(1), 32-48. Available at http://e-flt.nus.edu.sg/v7n12010/gomez.pdf

Gorard, S. (2014). A proposal for judging the trustworthiness of research findings. *Radical Statistics, 110*, 47-59.

Gorard, S. (2016). Damaging real lives through obstinacy: Re-emphasising why significance testing is wrong. *Sociological Research On-line*, *21*(1).

Gorard S., See, B. H., & Siddiqui, N. (2017). *The Trials of evidence-based education: The promises, opportunities and problems of trials in education*. London: Routledge.

Halpern, D. F. (1993). Assessing the effectiveness of critical thinking instruction. *The Journal of General Education, 42*(4), 238-254.

Halpern, D. F. (2014). Thought and Knowledge: An Introduction to Critical Thinking (5th ed). NY: Psychology Press.

Huber, C. R., & Kuncel, N. R. (2016). Does college teach critical thinking? A meta-analysis. *Review of Educational Research, 86*(2), 431-468. doi:10.3102/0034654315605917

Hurte, V. J. (2004). *A comparison of the scaffolding approach and the Cognitive Enrichment Advantage approach in enhancing critical thinking skills in first-year university freshman* (Ph.D.). The University of Tennessee. Retrieved from ProQuest Dissertations & Theses Global. (305133385). Available at https://trace.tennessee.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=6263&context=utk_graddiss

Iraji, H. R., Enayat, M. J., & Momeni, M. (2016). The effects of self- and peer-assessment on Iranian EFL learners' argumentative writing performance. *Theory and Practice in Language Studies*, *6*(4), 716-722.

Jafari, Z., & Yavari, S. (2014). The impact of conferencing on EFL learners' critical thinking. *Modern Journal of Language Teaching Methods, 4*(4), 153-156. Available at https://www.questia.com/library/journal/1P3-4014699921/the-impact-of-conferencing-on-efl-learners-critical

Jafari, Z., Yavari, S., & Ahmadi, S. D. (2015). The impact of self-assessment on EFL learners' critical thinking. *Modern Journal of Language Teaching Methods*, *5*(1), 145-149. Available at https://www.questia.com/library/journal/1P3-4017267101/the-impact-of-self-assessment-on-efl-learners-critical

Kahrizi, P., M.A., Farahian, M., & Rajabi, S. (2014). The impact of self-assessment on self-regulation and critical thinking of EFL learners. *Modern Journal of Language Teaching Methods, 4*(1), 353-370. Available at

https://www.questia.com/library/journal/1P3-3439447871/the-impact-of-self-assessment-on-self-regulation-and

Khamkhong, S. (2018). Developing English L2 critical reading and thinking skills through the Pisa Reading Literacy Assessment Framework: A case study of Thai EFL learners. *The Southeast Asian Journal of English Language Studies, 24*(3), 83-94. http://doi.org/10.17576/3L-2018-2403-07

Khatib, M., & Alizadeh, I. (2012). Critical thinking skills through literary and non-literary texts in English classes. *International Journal of Linguistics, 4*(4), 563-580.

Khatib, M., & Janpour, J. M. (2012). Literary Texts and Critical Thinking. *Advances in English Linguistics, 1*(2), 30-36. Available at http://www.macrothink.org/journal/index.php/ijl/article/view/2928

Khatib, M., Marefat, F., & Ahmadi, M. (2012). Enhancing critical thinking abilities in EFL classrooms: Through written and audiotaped dialogue journals. *Humanity & Social Sciences Journal, 7*(1), 33-45. Available at https://www.semanticscholar.org/paper/Enhancing-Critical-Thinking-Abilities-in-EFL-%3A-and-Khatib-Marefat/caff1ae335f1090cb19bc4e1ff76ea49e4f86b64

Khodadady, E., & Ghanizadeh, A. (2011). The impact of concept mapping on EFL learners' critical thinking ability. *English Language Teaching, 4*(4), 49-60.

Kline R. B. (2004). *Beyond significance testing. reforming data analysis methods in behavioral research*. Washington, DC: APA.

Kusumoto, Y. (2018). Enhancing critical thinking through active learning. *Language Learning in Higher Education, 8*(1), 45-63. doi:10.1515/cercles-2018-0003

Manning, W. J. H. (1997). *The relationship between critical thinking and attitudes toward reading of the community college student enrolled in a Critical Reading course at Roane State Community College* (Ed.D.). The University of Tennessee, Ann Arbor. Retrieved from ProQuest Dissertations & Theses Global. (304391102). Retrieved from https://search_proquest_com.ezphost.dur.ac.uk/pqdtglobal/docview/304391102/AAB3765F875F42CDPQ/1?accountid=14533

Marin, L. M., & Halpern, D. F. (2011). Pedagogy for developing critical thinking in adolescents: Explicit instruction produces greatest gains. *Thinking Skills and Creativity, 6*(1), 1-13.

Mazer, J. P., Hunt, S. K., & Kuznekoff, J. H. (2007). Revising general education: Assessing a critical thinking instructional model in the basic communication course. *The Journal*

of General Education, 56(3-4), 173–199.

McCarthy-Tucker, S. N. (1995). *Teaching reality-based formal logic to adolescents to improve critical thinking skills* (Ph.D.). Arizona State University. Retrieved from ProQuest Dissertations & Theses Global. (304174747). Retrieved from https://search_proquest_com.ezphost.dur.ac.uk/pqdtglobal/docview/304174747/2C402D3ED7946EFPQ/1?accountid=14533

McMillan, J. (1987). Enhancing college student's critical thinking: A review of studies. *Research in Higher Education, 26*(1), 3-29.

McPeck, J. E. (1984). Stalking beasts, but swatting flies: The teaching of critical thinking. *Canadian Journal of Education, 9*(1), 28-44.

Mitchell, R., Myles, F., Johnston, B., & Ford, P. (2003, May 23). Criticality and the 'Key Skills' Agenda in Undergraduate Linguistics. University of Southampton. Notes of talk given at subject Centre for Languages, Linguistics and Area Studies Seminar: 'Key Skills Linguistics', CILT; London.

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., Elbourne, D., Egger, M., & Altman, D. G. (2010). CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *Journal of Clinical Epidemiology, 63*(8), 1-28.

Moore, R. A. (1995). *The relationship between critical thinking, global English proficiency, writing, and academic development for 60 Malaysian second language learners* (Ph.D.). Indiana University. Retrieved from ProQuest Dissertations & Theses Global. (304200096). Retrieved from https://search_proquest_com.ezphost.dur.ac.uk/pqdtglobal/docview/304200096/9D31795A0B8D4EABPQ/1?accountid=14533

Moore, T. (2011). Critical thinking and disciplinary thinking: A continuing debate. *Higher Education Research and Development, 30*(3), 261–74.

Moore, T. (2014). Wittgenstein, Williams and the terminologies of higher education: A case study of the term 'critical'. *Journal of Academic Language & Learning, 8*(1), A95-A108.

Niu, L., Behar-Horenstein, L. S., & Garvan, C. W. (2013). Do instructional interventions influence college students' critical thinking skills? A meta-analysis. Educational Research Review, 9, 114-128. doi:10.1016/j.edurev.2012.12.002

Norris, S. P. (1985). Synthesis of research on critical thinking. *Educational Leadership, 42* (8), 40-45.

Pally, M. (2001). Skills development in "sustained" content-based curricula: Case studies analytical/critical thinking and academic writing. *Language and Education, 15*(4), 279-305.

Pashangzadeh, A., Ahmadian, M., & Yazdani, H. (2016). From narativity to criticality: Developing EFL learners' critical thinking skills through short narratives/stories reading. *Education and Linguistics Research, 2*(1), 98-119.

Paul, R., & Elder, L. (2001). *Critical thinking: Tools for taking charge of your learning and your life*. Upper Saddle River, NJ: Prentice Hall.

Paul, R., Elder, L, & Bartell, T. (1997). California teacher preparation for instruction in critical thinking: Research findings and policy recommendations. Sacramento, CA: California Commission on Teacher Credentialing. Retrieved from http://www.eric.ed.gov/contentdelivery/servlet/ERICServlet?accno=ED437379

Paul, R. (2004). The state of critical thinking today. *The Critical Thinking Community*. Retrieved from http://www.criticalthinking.org/pages/the-state-of-critical-thinking-today/523

Perezgonzalez, J. D. (2015). The meaning of significance in data testing. *Frontiers in Psychology, 6*(1293), 1-4 doi: 10.3389/fpsyg.2015.01293

Pithers, R. T., & Soden, R. (2000). Critical thinking in education: A review. *Educational Research*, 42(3), 237-249.

Ramanathan, V. & Kaplan, R. B. (1996). Some problematic "channels" in the teaching of critical thinking in current L1 composition textbooks: Implications for L2 student-writers. *Issues in Applied Linguistics, 7*(2), 225-249. Retrieved from http://escholarship.org/uc/item/8bn658q0

Rashtchi, M. (2007). A pathway toward critical thinking through cooperative writing in an English college course in Iran. *The Near and Middle Eastern Journal of Research in Education, 2*(1), 1-11.

Renaud, R. D., & Murray, H. G. (2008). Comparison of a subject-specific and a general measure of critical thinking. *Thinking Skills and Creativity, 3*(2), 85-93

Richmond, J. E. (2007). Bringing critical thinking to the education of developing country professionals. *International Education Journal, 8*(1), 1-29.

Robson, C. (2014). *Real world research: A resource for users of social research methods in applied settings*. (3rd ed.). Oxford, UK: Wiley-Blackwell.

Ruff, L. G. (2005). *The development of critical thinking skills and dispositions in first-year college students: Infusing critical thinking instruction into a first-year transitions course* (Ph.D.). University of Maryland, College Park. Retrieved from ProQuest

Dissertations & Theses Global. (304993178). Retrieved from https://search_proquest_com.ezphost.dur.ac.uk/pqdtglobal/docview/304993178/1ACEBA1406D84F5DPQ/1?accountid=14533

Salmani Nodoushan, M. A. (2016). Working on the 'write' path: Improving EFL students' argumentative-writing performance through L1-mediated structural cognitive modification. *International Journal of Language Studies, 10*(4), 131-152.

Shaarawy, H. Y. (2014). The effect of journal writing on students' cognitive critical thinking skills: A quasi-experimental research on an English as a foreign language (EFL) undergraduate classroom in Egypt. *International Journal of Higher Education, 3*(4), 120-128.

Sadler, T.D. (2006). Promoting discourse and argumentation in science teacher education." *Journal of Science Teacher Education, 17*(4), 23-346.

Schafersman, S. D. (1991, January). An introduction to critical thinking. *The Free Inquiry*. Retrieved from http://www.freeinquiry.com

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Boston, MA: Houghton Mifflin.

Shim, W., & Walczak, K. K. (2012). The impact of faculty teaching practices on the development of students' critical thinking skills. *International Journal of Teaching and Learning in Higher Education, 24*, 16–30.

See, B.H. (2016). An investigation into the teaching and learning of argumentation in first year undergraduate courses: A pilot study. *British Journal of Education, Society and Behavioural Science, 18*(4), 1-25.

Slavin, B. (2018, May 10). Effect sizes and the 10-foot man [Blog post]. Available at https://robertslavinsblog.wordpress.com/2018/05/10/effect-sizes-and-the-10-foot-man/

Sokol, A., Oget, D., Sonntag, M., & Khomenko, N. (2008). The development of inventive thinking skills in the upper secondary Language classroom. *Thinking Skills and Creativity, 3*(1), 34–46.

Ten Dam, G., & Volman, M. (2004). Critical thinking as a citizenship competence: Teaching strategies. *Learning and Instruction, 14*(4), 359-379.

Tiruneh, D. T., Verburgh, A., & Elen, J. (2014). Effectiveness of critical thinking instruction in higher education: A systematic review of intervention studies. *Higher Education Studies, 4*(1), 1-17.

Torgerson, C. J., Andrews, R. J., Robinson, A. M., & See, B. H. (2006). A systematic review

of effective methods and strategies for improving argumentation skills in undergraduate students in Higher Education. York: The Higher Education Academy.

Trafimow, D., & Rice, S. (2009). A test of the null hypothesis significance testing procedure correlation argument. *The Journal of General Psychology,136*, 261–269. doi: 10.3200/GENP.136.3.261-270

Tous, M. D., & Haghighi, S. (2016). Developing critical thinking with debate: Evidence from Iranian male and female students. *Informal Logic*, *36*(1), 64-82.

Tous, M. D., Tahriri, A., & Haghighi, S. (2015). The effect of instructing critical thinking through debate on male and female EFL learners' reading comprehension. *Journal of the Scholarship of Teaching and Learning*, *15*(4), 21-40.

Turuk K. M. C. (2011). *Developing critical thinking skills through integrative teaching of reading and writing in the L2 writing classroom* (Ph.D.). University of Newcastle Upon Tyne. Retrieved from ProQuest Dissertations & Theses: UK & Ireland. (1124082835). Retrieved from https://search_proquest_com.ezphost.dur.ac.uk/pqdtglobal/docview/1780169130/18E67C4F5FE346F8PQ/1?accountid=14533

Vacha-Haase, T. & Thompson B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology*, *51*(4), 473–481.

Yang, Y.-T. C., & Gamble, J.. (2013). Effective and practical critical thinking-enhanced EFL instruction. *ELT Journal: English Language Teachers Journal*, *67*(4), 398–412.

Zelizer, D. A. (2013). *Critical thinking: Comparing instructional methodologies in a senior-year learning community* (Ph.D.). Capella University. Retrieved from ProQuest Dissertations & Theses Global. (1318600938). Retrieved from https://search_proquest_com.ezphost.dur.ac.uk/pqdtglobal/docview/1318600938/FAEB73E198E64BA4PQ/1?accountid=14533

Appendix - Data extraction table

Studies with 2* rating

Author(s) + Year + Country + Database	Aim	Teaching strategy	Research design as stated by researcher(s)	Sample size + Instrumentation	Level + Age group + Duration of intervention	Major findings + Outcome	Major limitations mentioned by the author(s)	Quality judgment based on the "sieve" (see Section 2.5)
Gomez (2010) Colombia Handsearch	To examine the effect of structured reading lessons on the development of students' critical thinking skills	CT skills: analysis, application, evaluation and synthesis of information	Experimental (experimental and control) Pre-test post-test	83 students (43 in experimental and 40 in control – 8 classes) Pre-test and post-test (California Critical Thinking Test – Spanish version)	First-year university students (18 to 23 years old) 1 semester (15 weeks -2 sessions per week – 2 hours each)	No significant difference in scores between control groups and experimental groups *No effect*	Short duration of the intervention Inadequacy of the test to the context or culture	Strong design (random assignment of stds to 8 groups) 4* Attrition rate 0% in experimental group and 18% in control group due to schedule change (drops to 2*) **Low 2***
Kusumoto (2018) Japan Web of Science	To examine the effect of active learning	Active learning	Quasi-experimental (two groups)	162 students Cornell Critical Thinking Test (CCTT) Level Z as pre-test and post-test	University students 2 semesters	Experimental group showed improvement *No effect*	Lack of control over other English courses that students were taking	No randomization 3* No control over other English courses that students were taking (drops to 2.5*)

								Missing scores were eliminated from the analysis (drops to 2*) **Low 2***
Mazer, Hunt, & Kuznekoff (2007) U.S. JSTOR	To examine the effectiveness of a critical thinking instructional model in a communication course	Critical thinking skills: defining arguments, evaluating sources, identifying fallacies	Not specified by the researchers (random assignment of 18 clusters)	324 students (random assignment classes - 169 students in the control group and 155 students in the experimental group) Pre-test and post-test (researcher-developed with Kuder-Richardson, KR-20, reliability estimates of .84 for the pretest and .85 for the posttest)	University students (18 to 26 years old) (43% male and 57% female in experimental - 42% male and 58% female in control) 1 semester (16 weeks)	Students in the experimental group did better on the CT test than students in the control group *Positive effect*	A longer period of instruction needed No comparison of students' final course grades with their CT grades	Randomization of 18 clusters 4* Researcher-developed test (drops to 2*) Attrition was not reported (drops to 1*) **Low 2***
McCarthy-Tucker (1995) U.S. ASSIA	To examine whether instruction in formal logic can improve students' performance on both standardized and content-specific assessment of critical thinking, along with increased self-perception of critical thinking	Instruction in logic	Quasi-experimental (non-equivalent group design – pre-test post-test – cluster randomization)	189 students (9 sections) (62 students in the experimental group - 57 students in the second experimental group – 70 students in the control group) Pre-tests and post-tests: Raven's Standard Progressive Matrices (RSPM) +	Freshman and sophomore high school students (96 male students and 93 female students) 8 months	Experimental group outperformed control group on assessment of logic, assessment of thinking ability, and self-ratings of thinking skills *Positive effect*	Various threats to the internal validity of the study	Random assignment of 9 clusters 4* Attrition rate 38.8% (drops to 1*) **Low 2***

				Test of Logical Thinking (TOLT), Content-Specific + Test of Logic (CSTL) researcher-developed				
Salmani Nodoushan (2016) Iran ERIC	To investigate whether L1 mediated learning that aims at enhancing students' critical thinking skills would improve their argumentative writing	General critical thinking skills	Experimental with pre-test and post-test	894 students Argumentative essay The Cornell Critical Thinking Test, Form Z	3 weeks (three 2-hour sessions per week)	No difference in scores in essay writing between experimental and control Students in the experimental scored much higher than the control on the Cornell Critical Thinking Test – with students in higher proficiency levels scoring higher than those in low proficiency level *Positive effect*	None	Randomization of clusters to experimental and control groups – big number of cases 4* Not clear how many clusters were formed and who taught them 3* Two raters graded the essays but they were not blinded 3* Evidence of teaching the experimental group to the test (drops to 2*) **Low 2***

Studies with rating of 1.5*

Author(s) + Year + Country + Database	Aim	Teaching strategy	Research design as stated by researcher(s)	Sample size + Instrumentation	Level + Age group + Duration of intervention	Major findings + Outcome /reported effects	Major limitations mentioned by the author(s)	Quality judgment based on the "sieve" (see Section 2.5)
Tous, Tahriri, & Haghighi (2015) Iran ASSIA	To examine the effect of debate training on male and female reading comprehension	Debate	Experimental (2 groups – pre- and post-test design)	88 students (random assignment - 44 in 2 experimental groups and 44 in 2 control groups) Pre-test and post-test (Read Theory Critical Reading Comprehension Test + California Critical Thinking Skills Test – Persian version)	High-school students 1 month and a half	Debate has a statistically significant effect on students' reading comprehension No difference between males and females *Positive effect*	Duration of the study	Random assignment of stds to groups 4* Small number of cases (drops to 2*) Intervention is of short duration (1 month and a half) - short lapse between pre- and post-test (drops to 1.5*) **1.5***

Studies with rating of 1*

Author(s) + Year + Country + Database	Aim	Teaching strategy	Research design as stated by researcher(s)	Sample size + Instrumentation	Level + Age group + Duration of intervention	Major findings + Outcome	Major limitations mentioned by the author(s)	Quality judgment based on the "sieve" (see Section 2.5)
Akbari, Seifoori, & Ahour (2017) Iran Web of Science	To examine the effect of critical reading instruction on students' CT level	Critical reading skills like inferences, implications, probability	Random assignment of two intact classes	50 students Writing composition	Postgraduate students majoring in English (21 to 45 years old) 11 sessions of a 16 session course, each session lasting 90 minutes	Explicit CT awareness-raising is effective in enhancing experimental students' argumentative writing *Positive effect*	Short duration of the intervention	Random assignment of only 2 intact classes 3* Small number of cases – 2 clusters with 50 stds (drops to 2*) Essays were scored by 2 raters but they were not blinded + attrition was not reported (drops to 1*) **Extremely low 1***
Daud, Gilmore & Mayo (2013) Malaysia	To examine the usefulness of peer review, self-evaluation and peer evaluation on the development of students' critical	Peer review, self-evaluation and peer evaluation	Quasi-experimental (non-equivalent pre-test post-test design – 4 intact groups – 3 experimental	99 students Pre-test and post-test	Tertiary level university students 7 weeks	The peer-review group scored higher than other groups *Positive effect*	Time constraint for the peer evaluation group as there were more	No randomization 3* Short duration between pre-

Handsearch	thinking skills and writing ability To examine if a correlation exists between students' critical thinking skills and academic writing ability		groups and 1 control group)	(Cornell Critical Thinking Test - Level X) A final term paper			activities to be covered	and post-test may result in familiarity of stds with post-test (drops to 2*) No reporting of attrition (drops to 1*) **Extremely low 1***
Davidson & Dunham (1997) Japan Handsearch	To examine whether training in critical thinking enhances EFL learners' critical thinking level To test the suitability of a CT test developed by native speakers on non-native speakers	CT skills: logical fallacies, source credibility, inductive reasoning, informal deductive logic, and assumption-identification	Quasi-experimental (two-group post-test design)	36 students (17 experimental and 19 control) Post-test (Ennis-Weir Critical Thinking Essay Test)	First-year college students 1 year (13 hours of English per week) Some class hours (number not clear) lost due to an earthquake	Students in the experimental group outperformed those in the control group *Positive effect*	None	Not clear what the researcher means by "semi-lottery" randomization 3* Control group consisted of volunteers so they maybe they did not take the post-test seriously because it does not affect them in any way (drops to 2*) Very small number of cases (drops to 1*) **Extremely low 1***

Dong (2017) China Web of Science	To examine the effect of CT instruction on students' CT level	CT skills to guide writing	Experimental (two clusters) Pre-test post-test	44 students (22 in experimental and 22 in control) Essay	English major sophomore (22 years old) One semester	Improvement of CT level of the experimental group *Positive effect*	None	Random assignment of only 2 intact classes 3* Small number of cases – 2 clusters with 44 stds (drops to 2*) Essays were scored by 2 raters but they were not blinded + attrition was not reported (drops to 1*) **Extremely low 1***
Fatemi (n.d.) Iran Handsearch	To examine whether critical thinking skills can be taught to students by exposing them to literary narratives and "The Awareness of Consequences Technique"	Narrative texts	Quasi-experimental (pretest–posttest intact group design)	105 students (58 in experimental and 47 in control) Watson- Glaser Critical Thinking Appraisal (WGCTA) – Form A (Persian version)	EFL university students in their second semester (average age of 20) 1 semester (17 weeks – 2 sessions per week)	A significant improvement in critical thinking skills was shown in the experimental group *Positive effect*	None	Very weak design for RQ – unbalanced groups (no randomization) 1* **Extremely low 1***
Ghabanchi & Behrooznia (2014) Iran	To examine the impact of brainstorming on students' reading	Brainstormi ng	Experimental (intact group design – pre-test post-test)	54 students (25 in experimental and 29 in control)	University students in a reading course (30 females and 24 males)	Scores on the post-test show that brainstorming had a significant	None	No randomization – 2 intact groups) 3*

Handsearch	comprehension and critical thinking			Pre-test and post-test (the reading section of the TOEFL - and Watson-Glaser Critical Thinking Appraisal – Persian version)	16 sessions (90 minutess each)	effect on reading comprehension ability and critical thinking *Positive effect*		Small sample size (drops to 2*) attrition not reported (drops to 1) **Extremely low 1***
Jafari & Yavari (2014) Iran ASSIA	To investigate the effect of conferencing on students' critical thinking	Conferencing	Not specified by authors (2 groups with pre- and post-test design)	60 students (random assignment to 30 in experimental and 30 control) Pre-test and post-test (The Watson-Glaser test- Form A - Persian version)	Elementary adult EFL students 1 semester	The experimental group outperformed the control group *Positive effect*	None	Random assignment to groups 4* Small number of cases (drops to 3*) No clear description of what the treatment consisted of (drops to 1*) **Extremely low 1***
Jafari, Yavari, & Ahmadi (2015) Iran ASSIA	To investigate the effect of self-assessment on students' critical thinking and language proficiency	Self-assessment	Not specified by authors (2 groups with pre- and post-test design)	50 students (random assignment to 25 in experimental and 25 in control) Pre-test and post-test (The Watson-Glaser test- Form A - Persian version)	Intermediate adult learners 24 sessions	The experimental group outperformed the control group on both the critical thinking test and the English test *Positive effect*	None	Random assignment to groups 4* Small number of cases (drops to 3*) No clear description of what the

								treatment consisted of (drops to 1*) **Extremely low 1***
Kahrizi, Farahian, & Rajabi (2014) Iran ASSIA	To investigate the effect of self-assessment on students' self-regulation and critical thinking	Self-assessment	Not specified by authors (2 groups with pre- and post-test design)	40 students (random assignment to 20 in experimental and 20 in control) Pre-test and post-test (The California Critical Thinking Skills Test)	EFL learners from 3 different language schools (18 to 23 years old) 6 weeks	The experimental group made a significant gain in critical thinking *Positive effect*	None	Randomization is not clearly described 4* Small number of cases (drops to 3*) Short duration between pre- and post-test may result in familiarity of stds with post-test (drops to 2*) Attrition rate was not reported (drops to 1* **Extremely low 1***
Khatib & Alizadeh (2012) Iran	To examine the effect of using literary texts on students' critical thinking skills	Literary texts	Not specified by the authors (experimental and control groups– pre-test and post-test)	34 students (17 in experimental and 17 in control) Pre-test and post-test (Watson-Glaser Critical	Advanced language learners at a private language institute	Although both groups showed development in critical thinking and reading comprehension, the experimental	Not an equal number of males and females in the two groups	Randomization is not clearly described 3* Very small number of

Handsearch	To examine the effect of teaching critical thinking skills regardless of material			Thinking Appraisal - WGCTA and a test of reading comprehension - The 2005 TOFEL Test)	Twice a week – 70 days	group outperformed the control group. *Positive effect*		cases (drops to 1*) **Extremely low 1***
Khatib, Marefat, & Ahmadi (2012) Iran Handsearch	To examine the effect of keeping audiotaped and written dialogue journals on EFL students' critical thinking	Journal writing	Quasi-experimental (intact groups based on oral and written placement tests)	33 students (19 in the 1st experimental; 9 in the 2nd experimental; 12 in the control) Pre-test and post-test (Watson-Glaser Critical Thinking Appraisal - Form A – written in the Farsi language)	Female advanced EFL learners at an institute (19 to 33 years old) 1 semester (45 days – 6 hours per week)	Students in the two experimental groups outperformed their counterparts in the control group No difference in performance between the two experimental groups *Positive effect*	Small sample size	No randomization 3* Very small sample size (drops to 1*) **Extremely low 1***
Khodadady & Ghanizadeh (2011) Iran Handsearch	To examine whether concept mapping used as a post-reading strategy had an effect on EFL students' critical thinking ability	Concept mapping	Not specified by the authors (pre-test post-test intact group design)	36 students (18 in experimental and 18 in control) Pre-test and post-test (Watson-Glaser Critical Thinking Appraisal – Form A)	Intermediate and advanced EFL learners (31 females and 5 males) in a language center 22 two-hour sessions	Students in the experimental group outperformed those in the control group *Positive effect*	The sample is not representative in terms of age and gender	Randomization of stds to two groups 4* Very small number of cases (drops to 1*) **Extremely low 1***
Rashtchi (2007) Iran Handsearch	To investigate whether collaborative writing enhances	Cooperative writing	Not specified by the author (random assignment to two groups)	74 students (38 in experimental and 36 in control)	English translation university students (20	Students in the experimental group outperformed	None	Random assignment of stds to 2 groups 4*

	critical thinking skills			Pre-test and post-test (The Watson-Glaser Critical Thinking Appraisal, Form A (WGCTQ) + an essay graded by two raters)	males and 70 females) 1 semester (14 sessions)	those in the control group *Positive effect*		Medium number of cases (drops to 3*) Not clear whether the raters of the writing test were blinded (drops to 2*) Attrition rate is not reported (drops to 1) **Extremely low 1***
Ruff (2005) U.S. ASSIA	To examine whether students who are enrolled in a transitions course in which critical thinking skills and dispositions are taught do better than students who are enrolled in the same course but do not receive instruction in critical thinking	Critical thinking skills	Quasi-experimental (pre-test and post-test with no randomization)	39 students (20 students in the experimental group and 19 in the control group) Pre-test and post-test The California Critical Thinking Skills Test (CCTST) and the California Critical Thinking Dispositions Inventory (CCTDI)	University students 1 semester	Students in the experimental group scored higher than students in the control group regardless of gender *Positive effect*	No single agreed-upon definition for CT and no single agreed-upon strategy for teaching and testing CT in the literature Small non-random sample	No randomization 3* Small number of cases (drops to 2*) Diffusion of treatment: stds were taught by the same teacher (drops to 1) **Extremely low 1***

Sokol, Oget, Sonntag, & Khomenko (2008) Latvia and France ASSIA	To explore the effect of integrating inventive thinking skills instruction (The Thinking Approach) in foreign language teaching	Thinking Approach (TA) to language teaching and learning (Scaffolding)	Quasi-experimental (pre-test and post-test - no randomization)	81 students (54 students in the experimental group and 27 in the control group) pre-test and post-test (researcher-developed test)	Upper secondary students (15 to 16 years old) 1 academic year	Students in the experimental group showed a significant increase in thinking skills compared to the control group *Positive effect*	Contact hours were not the same for the groups Different level of language proficiency between control group and experimental group Students in the experimental group took the test more seriously	No randomization 3* Unbalanced groups in terms of language competency – from two schools (drops to 2*) Other threats: experimental stds had 5 hours of instruction per week while control stds had 3 hours per week + researcher-developed marked by the researcher who was not blinded + attrition is not reported (drops to 1*) **Extremely low 1***
Yang & Gamble (2013) China PsychINFO	To investigate if CT integration in the EFL curriculum can result in higher English proficiency	Argumentative writing and debating	Experimental (two intact groups – pre- and post-test)	68 students (random assignment of intact classes: 31 in experimental and 37 in control)	Freshman English Reading and Listening students (EFL learners)	Students in the experimental group did better on the post-test in terms of language proficiency,	None	Random assignment of only 2 intact classes 3*

	and higher level of critical thinking			The General English Proficiency Test (high-intermediate level) (reading and listening sections) An essay scored with the Holistic Critical Thinking Scoring Rubric A content-based achievement test (researcher-developed)	1 semester (8 weeks)	critical thinking, and academic achievement than students in the control group *Positive effect*		Small number of cases – 2 clusters with 68 stds (drops to 2*) Essays were scored by 2 raters but they were not blinded + attrition was not reported (drops to 1*) **Extremely low 1***
Zelizer (2013) U.S. ASSIA	To compare the effect of a mixed instructional approach (experimental) to critical thinking compared to an immersion approach (control) on students' development of critical thinking To analyse the extent to which students can transfer critical thinking skills learned in one course to another	Mixed instructiona l approach to teaching critical thinking	Quasi-experimental (nonequivalent group design – no randomization - convenience sampling pre-post-test design)	171 students (experimental group = 92 - control group = 79 – 4 classes) Pre-test and post-test (The Watson Glaser Critical Thinking Appraisal – Forms A and B)	Senior-year university students (19 to 47 years old) 1 semester	No difference in pre-test and post-test scores between experimental and control groups *Negative effect*	The results of a convenience sample cannot be generalizable	No randomization – 4 intact classes 3* The intervention consisted of material taken from the test – threat of teaching to the test (drops to 2*) Other weaknesses: Unbalanced dropout + Exclusion of

	course in the same semester							withdrawn participants from the analysis + Same teacher teaching all 4 classes which might have resulted in diffusion of treatment (drops to 1*) **Extremely low 1***

Studies with the rating of 0

Author(s) + Year + Country + Database	**Aim**	**Teaching strategy**	**Research design as stated by researcher(s)**	**Sample size + Instrumentation**	**Level + Age group + Duration of intervention**	**Major findings + Outcome**	**Major limitations mentioned by the author(s)**	**Quality judgment based on the "sieve" (see Section 2.5)**
Arslan & Yildiz (2012) Turkey Handsearch	To examine the application of a literature-based critical thinking programme on students' critical thinking skills To examine the beliefs of both students' and	Literature-based critical thinking program	Quasi-experimental (one-group pre-test post-test design)	34 students Pre-test and post-test (Cornell Critical Thinking Test - Level Z)	Undergraduate fourth-year university students (31 females and 3 males) 7 weeks (13 sessions- 39 hours)	Students scored higher on the post-test than they did on the pre-test *Positive effect*	None	Very weak design for RQ – no comparison group 1* More weaknesses: no reporting of attrition +

	teachers' about literature instruction							short duration between pre- and post-test (same test) may result in familiarity of stds with post-test (drops to 0) **Rating 0**
Chason, Loyet, Sorenson, & Stoops (2017) It is not clear where the study was conducted as this was not reported, but it is deduced from the participants that the study took place in Saudi Arabia	To examine whether the TBSIR (topic, bridge, support, interpretation, return) framework has an effect on students' paragraph writing	General critical thinking skills: the TBSIR (topic, bridge, support, interpretation, return) framework in paragraph writing	Pre-experimental	37 students	Intermediate to advanced students enrolled in an 8-week course	Students made progress with this approach *Positive effect*	No control group to compare results with	Very weak design for RQ: No control group 1* Attrition was not reported (drops to 0) Short duration of intervention – 8 weeks **Rating 0**
Fahim & Mirzaii (2013) Iran Handsearch	To examine the effect of dialogic critical thinking on the writing performance of students	Dialogic critical thinking tasks	Quasi-experimental (randomized clusters experimental and control)	43 students (4 classes - 21 in experimental and 22 in control) Pre-test and post-test (in-class 180-word four-paragraph argumentative essay – two different topics in pre and post)	Upper-intermediate EFL male learners at an institute (17 to 41 years old) 1 semester (5 weeks – 21 sessions)	Although both groups showed improvement in argumentative writing, the experimental group exhibited superior performance *Positive effect*	The study included only males The study was about written production and could not include oral production	Randomization of only 4 clusters – very small number of cases 2* No reporting of attrition (drops to 1* Researcher-developed test (possibility of teaching to the test) – no

								mention of blinding raters (drops to 0) **Rating 0**
Hurte (2004) U.S. ASSIA	To compare the effectiveness of two approaches (a Scaffolding approach and a modified, condensed version of the Cognitive Enrichment Advantage, CEA, approach) in enhancing students' critical thinking skills	Scaffolding approach and a student-centered approach	Quasi-experimental (Pre-test post-test comparison group design – matched pairs to 2 experimental groups – based on the WGCTA)	36 students (random assignment of matched pairs to 2 experimental groups with no control group) Pre-test and post-test (Watson-Glaser Critical Thinking Appraisal – Forms A & B and the critical thinking performance assessment)	First-year university Freshman students 1 semester (16 weeks) Intervention phase: five weekly 40-minute teaching sessions	No significant change in CT in the CEA group based on the two assessment tools No significant change in the Scaffolding group based on CT performance assessment and a decline based on the W-GCTA *Negative effect*	Short duration of the study Diffusion of treatment Lack of a control group Researcher acting as instructor	No comparison group 1* Very small number of cases (drops to 0) Short duration of intervention to result in any change – 5 sessions only (drops to 0) **Rating 0**
Iraji, Enayat, & Momeni (2016) Iran ASSIA	To examine the effect of self-assessment and peer-assessment on students' argumentative writing	Assessment techniques	Not specified by authors (Pre-test and post-test – 2 groups)	36 students (random assignment to experimental and control groups) Pre-test and post-test (an argumentative essay)	Intermediate EFL students (18 to 25 years old) Not stated	The experimental group outperformed the control group *Positive effect*	None	Random assignment to groups 4* Very small number of cases (drops to 1*) Other threats: no mentioning of number of raters and whether they were blinded + duration of intervention

								not mentioned + researcher developed test which might result in teaching to the test (drops to 0) **Rating 0**
Khamkhong (2018) Thailand Web of Science	To test the effectiveness of the PISA reading literacy framework on students' level of critical thinking	Literary texts: The PISA reading literacy framework	Pre-experimental	36 students	Third-year English majors 16 weeks	Students made progress with this approach *Positive effect*	None	Very weak design for RQ: No control group 1* Researcher-developed test (drops to 0) **Rating 0**
Khatib & Janpour (2012) Iran Handsearch	To investigate the effect of literary texts on the development of students' critical thinking	Literary texts	Experimental	30 students (15 students in experimental and 15 in control) Pre-test and post-test (Watson-Glaser Critical Thinking Appraisal questionnaire)	Advanced students (19 to 27 years old) 20 sessions	Students in the experimental group performed better in the post-test than students in the control group *Positive effect*	None	Students were matched and then randomized 3* Very small number of cases (drops to 1*) Attrition rate was not reported (drops to 0) No control over confounds – did the texts or

								the questions effect a change in students' critical thinking **Rating 0**
Manning (1997) U.S. ASSIA	To determine the relationship between students' critical thinking and their attitudes to reading To determine the effect of critical thinking instruction on students' critical thinking	Critical thinking skills: perceiving, classification, concept formation, identification patterns and relationships, and problem solving	Not specified (non-equivalent group design – no randomization)	31 students (15 in the experimental and 16 in the control taught by the same instructor) Pre-test and post-test (The Cornell Critical Thinking Test, Level X) Rhody Secondary Reading Attitude Assessment	University students 1 semester (5 weeks)	No significant correlation between attitude to reading and critical thinking in both the control and treatment groups A significant difference in critical thinking in pre-test and post-test scores in both groups (higher in control) *Negative effect*	Teacher taught both groups Short duration of the study Small sample size	Very weak design for RQ: no randomization 3* Completely unbalanced groups from 2 different campuses – researcher admits that the 2 groups are different (drops to 2*) Other weaknesses: very small number of cases + diffusion of treatment – same instructor teaching both groups +attrition rate was not reported (drops to 0)

								0
Moore (1995) U.S ASSIA	To examine the relationship between critical thinking skills and language proficiency, writing, and academic development To examine the effect of critical thinking instruction on students' scores on a CT test	Critical thinking skills: identifying issues, conclusions, reasons, assumptions, errors in reasoning, etc.	Pre-experimental (single group design - pre-test and post-test)	60 students Pre-test and post-test (The Ennis-Weir Critical Thinking Essay Test – essay form)	Pre-university students in a critical thinking course 1 semester (16 weeks)	Significant gains in critical thinking between pre-test and post-test Language proficiency has a significant relationship with CT Writing ability and academic development in English have no significant relationship with CT *Positive effect*	Small sample size No control group Students selected for the study are top-quality Malaysian students	Very weak design for RQ – no comparison group 1* Maturation threat: sample consisted of high-achievers who were selected to move from Malaysia to the U.S. – can't be sure if moving to the U.S or the intervention resulted in this growth (drops to 0) **Rating 0**
Pashangzadeh, Ahmadian, & Yazdani (2016) Iran Handsearch	To investigate the effect of narrative texts on students' critical thinking	Narratives	Not specified by authors (two intact groups – pre- and post-test)	54 students (27 in each group) Pre-test and post-test (California Critical Thinking Skills Test)	Undergraduate EFL learners majoring in translation 12 treatment sessions	Students in the experimental group outperformed those in the control group *Positive effect*	None	No randomization – 2 intact groups) 3* Small number of cases (drops to 2*) Not clear what the control

								group did (the non-narrative group) – it might be that they did not do anything useful in class (drops to 1*) Not clear who taught the two groups (drops to 0) Attrition rate not reported **Rating 0**
Shaarawy (2014) Egypt ASSIA	To examine the effect of weekly academic journal writing on students' critical thinking	Journal writing	Quasi-experiment (pre- and post-test)	23 students (16 in experimental and 7 in control) Pre- and post-test (researcher-developed based on Bloom's taxonomy)	First year university students in their 2nd semester 1 semester (7 weeks)	Students in the experimental group outperformed students in the control group *Positive effect*	Small sample size Short duration of intervention	No randomization 3* Very small number of cases (drops to 1*) Exclusion of participants who dropped from the final analysis of results instead of using intention-to-treat analysis (drops to 0)

								Researcher-developed test marked by the researcher who was not blinded **Rating 0**
Tous & Haghighi (2016) Iran Web of Science	To investigate whether there is any difference between males and females in critical thinking after instruction in debate	Debate	(1 group - pre-test and post-test)	88 students Pre-test and post (California Critical Thinking Skills Test – Form B - Persian version)	High school students (17 years old) 1 month	No difference between males and females *Negative effect*	Duration of the study	Poor reporting (not clear whether they were all placed in one group or split – if split, not clear whether groups consisted of both males and females and who taught the groups) 1* Intervention is of short duration (1 month) - short lapse between pre- and post-test so threat of stds becom ing familiar with the post-test (drops to 0) **Rating 0**

Turuk Kuek (2011) United Kingdom ASSIA	To find out if ESL students' reasoning and critical thinking as manifested in their writing improves as a result of an integrative approach to teaching reading and writing supported by collaboration and scaffolding	Critical thinking skills and collaboration : identification of author's viewpoint in a written text, the reason(s) offered to support the viewpoint, etc.	Experimental -Randomized controlled trial	20 students (randomly assigned to 11 in the experimental group and 9 in the control group taught by the same instructor) Pre-test and post-test (written composition test graded based on the following rubrics: Stapleton's (2001) model of assessing critical thinking in writing and Connor & Lauer's (1985) and Connor's (1990) scale of the persuasiveness of rational, credibility and affective appeals	First-year university students from the Faculty of Medicine at the Schools of Medicine and Nursing (17 to 34 years old) 12 weeks	Students in the experimental group scored much higher on their writing than those in the control group *Positive*	Students' weaknesses in the language had to be ignored in the scoring process The influence of reading on writing was investigated but the influence of writing on reading was not Short duration of the study	Random assignment of stds to groups 4* Very small number of cases (drops to 1*) Attrition rate was high – 27 did the pre-test – 47% + exclusion of scores of stds who dropped instead of using intention-to-treat analysis + researcher-developed test – 2 raters but not blinded (drops to 0) **Rating 0**