The effectiveness of debate pedagogy in ESL/EFL speaking skills: A systematic review

Xiaoying Wu

MSc in Applied Linguistics and Second Language Acquisition, 2023

DECLARATION BY THE CANDIDATE AS AUTHOR OF THE DISSERTATION

1. I understand that I am the owner of this dissertation and that the copyright rests with me unless I specifically transfer it to another person.

2. I allow the Department to deposit on my behalf a copy of this dissertation in the Oxford University Research Archive ('ORA') where it shall be freely available online for use in accordance with ORA's Terms and Conditions of Use [https://ora.ox.ac.uk/terms_of_use].

3. I understand that this dissertation should not contain material that can be used to personally identify individuals or specific groups of individuals (unless permission has been obtained from the individuals) and that such material should be removed before this dissertation is deposited in ORA.

4. I agree to be bound by the terms of the ORA Grant of Non-exclusive Licence [https://ora.ox.ac.uk/deposit_agreements] and I warrant that to the best of my knowledge, making my thesis available on the internet will not infringe copyright or any other rights of any other person or party, nor contain defamatory material.
5. I agree that my dissertation shall be available for download in ORA in accordance with paragraphs 2, 3 and 4 above.

Signed [an electronic signature is sufficient]:	Xiaoying Wu
Date:	17/08/2023

Abstract

This systematic review critically evaluated the role of classroom debates in enhancing the English-speaking skills of English as a second language or English as a foreign language (ESL/EFL) learners. 29 studies were included in this review, predominantly from Asia, followed by Europe and the Middle East, and they were conducted in secondary and tertiary settings. Heterogeneity was found across the studies regarding the research designs, debate process characteristics, and measurement instruments.

Overall, the included studies found positive impacts of debate pedagogy in developing ESL/EFL speaking skills. The reported outcomes in language aspects mainly encompassed the ability to construct and refute arguments, expand vocabulary, and enhance fluency. There were also observed positive impacts on affective dimensions such as increased confidence, reduced anxiety, and heightened motivation in second language (L2) speaking. Moreover, based on the evidence examined, this review identified elements to establish a debate-based teaching schedule that has the potential to facilitate instructors in incorporating debate pedagogy into their teaching practices and guide learners in practicing and refining their language skills through structured debate activities.

The trustworthiness appraisal, however, revealed that the overall methodological quality of included studies was poor. The identified research designs revealed numerous methodological issues, rendering it challenging to determine the direct impacts of debate pedagogy on ESL/EFL speaking skills and limiting the generalisability of the findings. These concerns primarily include: potential practice effects, namely the observed positive outcomes may be attributed to more speaking practice rather than the direct impacts of debate intervention; perceived effects rather than actual improvements based on self-reported data; the selection bias arising from the predominantly intermediate and advanced proficiency level participants, with less representation of beginner-level students, and the selection mainly from intact classes limit the generalisability of findings to broader populations; the potential influence of motivation based on voluntary participation in some studies; limited study locations, mostly being single-nation or single-institution; a lack of longitudinal studies and delayed post-tests to examine the sustained effects of debate intervention; and the shortage of robust randomised controlled trials (RCTs) to ensure that any group

differences are due to the intervention rather than other factors. Accordingly, future studies should adopt more robust research designs to address these concerns and ensure the benefits of debate pedagogy for learners of varying proficiency levels.

Acknowledgments

Studying at Oxford has been an incredibly enriching experience, and I am sincerely grateful for the assistance and encouragement that have paved the way for me to complete this dissertation.

First and foremost, I would like to thank my supervisor, Dr Hamish, for all the helpful advice, guidance, and support throughout my academic journey. Your constant encouragement has given me the confidence to overcome challenges and strive for excellence.

Thank you, Cate, for your support in statistics, data management, and analysis. Your advice and encouragement have been inspirational.

Thank you, Cathy, our librarian, for your assistance in the literature search, which helped me devise my search.

To Dr Jiang and Dr Zhang, and my dearest friends Anita and Wanqi, thank you for all the lovely conversations, intellectual and emotional support.

To the two feline companions of the St Hugh's College, Admiral Flapjack and Professor Biscuit. Your presence and moral support have made my life at Oxford more enjoyable.

My deepest gratitude to my parents, two elder brothers and other family members. Thank you for your endless love and support.

Table of contents

Abstract i
Acknowledgments iii
Table of contents iv
List of figures vii
List of tables vii
Abbreviations viii
Chapter 1. Introduction 1
1.1 Background and objectives 1
1.2 Outline of the thesis 2
Chapter 2. Background and rationale 4
2.1 Speaking skills in the ESL/EFL context 4
2.1.1 Components of speaking skills 4
2.1.2 Challenges of learning and teaching ESL/EFL speaking 6
2.1.3 Approaches to learning and teaching ESL/EFL speaking 7
2.2 Debate as an educational approach: theoretical grounds 8
2.2.1 Definition of debate 8
2.2.2 Development and procedure of debate 9
2.2.3 Output hypothesis and interaction hypothesis 15
2.3 Debate as an educational approach: pedagogical grounds 16
2.3.1 Language aspects of ESL/EFL speaking 17
2.3.2 Affective aspects of ESL/EFL speaking 19
2.3.3 Critiques of debate pedagogy 20
2.4 Gaps in the literature 22
2.5 Research questions 23
Chapter 3. Methods 24
3.1 Eligibility criteria 26
3.2 Information sources 29
3.3 Search strategy 31
3.4 Data management 32
3.5 Selection process 33
3.5.1 Title and abstract screening 33
3.5.2 Full-text screening 33
3.5.3 Ensuring screening reliability 33
3.6 Data extraction 34
3.7 Trustworthiness appraisal 35
3.8 Data synthesis 37
Chapter 4. Results 38

4.1 Results of the search and selection process 38
4.2 General characteristics 42
4.2.1 Publication details 42
4.2.2 Geographic and educational contexts 44
4.2.3 Research design 45
4.2.4 Sample characteristics 46
4.2.5 Intervention details 48
4.2.6 Methods of data collection 53
4.2.7 Reported outcomes 54
4.3 Trustworthiness appraisal of individual studies 56
4.3.1 Screening questions 59
4.3.2 Studies reporting qualitative data 59
4.3.3 Studies reporting quantitative data 59
4.3.4 Studies with a mixed methods design 61
4.3.5 Summary of methodological quality 61
4.4 The findings of included studies 62
4.4.1 Qualitative studies 62
4.4.2 Quantitative studies 62
4.4.3 Mixed methods studies 63
4.4.4 Effect sizes 64
Chapter 5. Discussion 70
5.1 Geographic and educational contexts 70
5.1.1 Location of studies 70
5.1.2 Educational contexts 71
5.2 Implementation of debate pedagogy 72
5.2.1 Pre-debate activity 72
5.2.2 Actual debate 73
5.2.3 Post-debate activity 74
5.2.4 Proposed debate implementation schedule 75
5.3 ESL/EFL speaking outcomes 79
5.3.1 Language aspects 79
5.3.2 Affective aspects 80
5.4 Methodological issues in research designs 82
5.4.1 Studies reporting qualitative data 82
5.4.2 Studies reporting quantitative data 83
5.4.3 Summary of methodological issues 86
5.5 Implications of this review 87
5.6 Limitations of this review 88
Chapter 6. Conclusion 90
References 92
Appendix A IDESR Protocol 107
Appendix B Preliminary screening process using ASReview 115
Appendix C Data extraction form 118

Appendix D Mixed Methods Appraisal Tool (MMAT, version 2018) 126
Appendix E Reporting the results of the MMAT (version 2018) 132
Appendix F List of references of included studies 134
Appendix G Basic information of included studies 137
Appendix H Sample characteristics 144
Appendix I Intervention details 148
Appendix J Discussed topics 165
Appendix K Topic word frequency 169
Appendix L Categorisation of outcomes in language aspects 173
Appendix M Narrative summary of included studies 178

List of figures

Figure 1 . Lincoln-Douglas debate format 11
Figure 2 . The public forum debate format 12
Figure 3 . Asian Parliamentary debate format 13
Figure 4 . British Parliamentary debate format 14
Figure 5 . Flow diagram of the screening process 39
Figure 6 . Publication year 43
Figure 7 . Publication type 43
Figure 8 . Research methodology 46
Figure 9 . Sample size 47
Figure 10 . Actual debate process 51
Figure 11 . Topic words frequency cloud 53
Figure 12 . Methods of data collection 53
Figure 13 . Frequency of reported outcome in language aspects 55
Figure 14 . Frequency of reported outcome in affective aspects 56

List of tables

Table 1 . TOEFL and IELTS speaking criteria 5
Table 2 . Eligibility criteria 26
Table 3 . List of databases 30
Table 4 . List of search terms 31
Table 5 . Study ID number 40
Table 6 . Location of studies and education phase 44
Table 7 . Results of the trustworthiness appraisal 58
Table 8 . Reported outcomes and overall results 65
Table 9 . Effect size 68
Table 10 . Proposal of debate-based learning and implementation schedule 77

Abbreviations

ASEAN: Association of Southeast Asian Nations

ASReview: The Active learning for Systematic Reviews

ACTEFL: American Council on the Teaching of Foreign Languages

CEFR: Common European Framework of Reference for Languages

EFL: English as a Foreign Language

ESOL: English for Speakers of Other Languages

EPPI: Evidence for Policy and Practice

ESL: English as a Second Language

FL: Foreign language

IELTS: International English Language Testing System

IDESR: International Database of Education Systematic Reviews

L1: First language

L2: Second language

MMAT: Mixed Methods Appraisal Tool

MM: Mixed Methods

PRISMA: Preferred Reporting Items for Systematic Review and Meta-Analysis

PROSPERO: International Prospective Register of Systematic Review database

QUAL: Qualitative

QUAN: Quantitative

RCTs: Randomised controlled trials

SOLOM: Student Oral Language Observation Matrix

TOEFL: Test of English as a Foreign Language

Chapter 1. Introduction

1.1 Background and objectives

The English language is widely recognised as a crucial global communication tool (Dearden, 2014; Rao, 2019b; Zhang & Zeegers, 2010). On a broader scale, English plays an influential role in diplomacy, cultural exchanges, international trade, and scientific and technological advancements (Iman, 2017). For individual learners of English as a Second Language (ESL) or English as a Foreign Language (EFL), proficiency in English – a lingua franca – serves an array of purposes, including overseas studies, access to academic resources, and enhancement of career opportunities (Rao, 2019a). As such, the ability to clearly express one's thoughts, converse, and speak in English is an essential skill widely sought by ESL/EFL learners (Schneider, 2020; Richards & Renandya, 2002).

However, the instruction and acquisition of ESL/EFL speaking skills face numerous obstacles (Abrar et al., 2018). In particular, in non-English speaking regions and countries, the lack of a vibrant ESL/EFL speaking community confines language learning and practice primarily in the classroom (Adem & Berkessa, 2022). This predicament is further exacerbated by the insufficient focus on ESL/EFL speaking practice and ineffective pedagogical methods in many educational settings, thereby restricting learners' opportunities to develop their speaking skills (Goh, 2017; Thornbury, 2005). Such constraints necessitate implementing meaningful pedagogical approaches to facilitate learners' development of speaking in the target language.

In light of these challenges, debate, as a pedagogical tool, holds the potential to effectively enhance learners' language skills, including speaking skills (El Majidi, 2022). Originating from ancient Greek educational practices, debate is characterised as a method of inquiry and advocacy that leads to reasoned judgement on a proposition (Freeley & Steinberg, 2013). This instructional method fosters critical thinking, interaction, idea exchange, and rapid cognitive processing, which are

indispensable for language acquisition and oral communication (Brown, 2009; Jodoi, 2023; O'Mahoney, 2015).

Whilst debate pedagogy is extensively applied within first language (L1) contexts, it has increasingly attracted attention in the second language (L2) field in recent decades (Zare & Othman, 2015). Studies have underscored the merits of participating in debates, with educators and learners perceiving improved overall language proficiency, particularly in speaking skills (e.g., Al-Mahrooqi & Tabakow, 2015; El Majidi et al., 2018; O'Mahoney, 2015). Recent empirical evidence lends further credence to the efficacy of debate in enhancing L2 speaking skills (e.g., El Majidi, 2021a; García-Sánchez, 2020; He, 2015). Furthermore, it has been reported that students' confidence improved whilst their language anxiety decreased after participating in classroom debates (e.g., Al-Tamimi, 2017; Rosas-Maldonado et al., 2020).

However, notwithstanding the consistently reported positive outcomes, the field is plagued by methodological shortcomings in research designs, making it difficult to draw any definitive conclusions regarding the direct impacts of debate pedagogy on L2 speaking development (see Section 5.4.3 for the summary of methodological issues). This dissertation adopts a systematic review methodology in order to comprehensively evaluate the research landscape on the effectiveness of classroom debate as an educational approach for developing ESL/EFL speaking skills on a global scale and to assess the robustness of these findings.

1.2 Outline of the thesis

Chapter 2 outlines the background and rationale of the study, highlighting the significance of developing speaking skills in the ESL/EFL context and positing classroom debate as a potential pedagogy for language development. Following this, Chapter 3 explains the methods applied in this study. Chapter 4 provides the findings derived from the studies identified. Subsequently, Chapter 5 delves into a discussion of these findings. This chapter also discusses implications for future research, L2

teaching and learning, as well as the limitations of this review. Finally, Chapter 6 presents conclusions of this study.

Chapter 2. Background and rationale

In this chapter, I establish the importance of speaking skills in ESL/EFL learning environments, considering their constituent parts and the challenges associated with learning and teaching. Based on this, I discuss classroom debate as a potentially effective educational approach to develop ESL/EFL learners' speaking skills, covering its theoretical and practical grounds. Finally, I identify gaps in the existing literature and establish the research questions for the present study.

2.1 Speaking skills in the ESL/EFL context

2.1.1 Components of speaking skills

Speaking is a dynamic and multifaceted process that involves creating and processing information to convey meaning (Burns & Joyce, 1997; Hughes & Reed, 2016). According to Levelt (1995, cited in Goh & Burns, 2012), the process of speaking has three interconnected stages, namely, conceptualisation (devising ideas), formulation (deciding on the words and structure to express these ideas), and articulation (physically producing the sounds), accompanied by an ever-present process of self-monitoring to ensure clarity and accuracy.

There are certain core elements when evaluating speaking skills. Commonly agreed key components include vocabulary, grammar, pronunciation, fluency, and comprehension (Lado, 1995; Brown & Nation, 1997; Gatbonton & Segalowitz, 2005). These elements have been incorporated into internationally recognised English language assessments, such as the Test of English as a Foreign Language (TOEFL) and the International English Language Testing System (IELTS), although with some variation in their evaluation of ESL/EFL speaking skills. Table 1 presents the speaking rubrics of these two assessments.

Table 1. TOEFL and IELTS speaking criteria

TOEFL[1]	IELTS[2]
General description (task fulfilment, intelligibility, and coherence of the speech)	Fluency and coherence (the ability to talk with normal levels of continuity, rate, and effort and to link ideas together coherently)
Language use (complexity and accuracy of grammar and vocabulary)	Lexical resource (range of vocabulary)
Topic development (coherence, elaboration, and clear links between ideas)	Grammatical range and accuracy (appropriate use of syntactic forms)
Delivery (pronunciation, flow, intonation, pacing)	Pronunciation (accurate use of phonological features to convey meaning)

Whilst these criteria have been criticised for their perceived ambiguity and subjectivity (Pearson, 2019), this section aims not to discuss these shortcomings but rather to understand these components to aid in relevant literature screening and data extraction concerning reported outcomes for this systematic review.

Collectively, speaking skills encompass a range of interconnected elements, including linguistic components such as vocabulary, grammar, pronunciation, and their utilisation in terms of fluency, accuracy, complexity, and coherence.

[1] https://www.ets.org/pdfs/toefl/toefl-ibt-speaking-rubrics.pdf

[2] https://www.ielts.org/-/media/pdfs/ielts-speaking-key-assessment-criteria.ashx

2.1.2 Challenges of learning and teaching ESL/EFL speaking

Speaking is a key component of English language instruction as many ESL/EFL learners consider the mastery of speaking skills a primary indicator of their success in language acquisition and the effectiveness of their English courses (Liao, 2009; Rao, 2019a; Richards, 2006). The value of English-speaking proficiency is not only in terms of personal success; it is also considered a vital asset in professional contexts, as many employers prioritise this skill during recruitment (Adem & Berkessa, 2022). Despite this recognition, speaking remains the most demanding among the four core language skills (reading, listening, speaking and writing), especially in a foreign language (FL) or L2 context (Jung, 1995; Rahimi & Quraishi, 2019).

The difficulties in learning and teaching ESL/EFL speaking could be attributed to its intrinsic complexity and external contextual factors (Adem & Berkessa, 2022). As discussed in Section 2.1.1, speaking is a complex process encompassing a range of sub-components. Moreover, the lack of an L2 speaking community often confines language practice to the classroom (Bueno et al., 2006). Compounding the issue, classroom practices often lack adequate time and emphasis on activities and tasks that foster speaking in students' L2 or FL (Zare & Othman, 2015).

Researchers express concerns over ESL/EFL students' speaking capacities in diverse contexts (e.g., Liu, 2015; Taddese, 2019; Wahyuningsih & Afandi, 2020). For instance, Noor and Tajik (2022) argued that traditional lecture-based courses in Afghanistan often limit students' opportunities for interaction and communication. El Majidi (2022) also claimed that speaking skills received the least attention compared to other skills in the teaching practices of Dutch secondary school English teachers, and the absence of effective oral teaching approaches undermines learners' efforts to fully develop their English communicative competence. Similarly, Asfaw et al. (2021) unearthed some factors hindering effective speaking instruction in Ethiopia, such as overlooked speaking lessons, ineffective classroom management, a lack of interactive

speaking activities, a fear of making errors, and an uneven level of participation across students.

One mentioned recurring issue mentioned above is the lack of effective educational approaches to foster L2 speaking skills. In response to this challenge, the following section briefly reviews different approaches to teaching speaking skills before proposing the adoption of classroom debates as a potentially effective pedagogy.

2.1.3 Approaches to learning and teaching ESL/EFL speaking

As discussed, classrooms are the primary environment for ESL/EFL students to actively learn and practice the language (Adem & Berkessa, 2022). Different approaches to teaching speaking skills have been explored, which, according to Goh and Burns (2012), can be broadly grouped into two principal methods, namely direct and indirect. The direct approach focuses on practicing and enhancing individual language forms, which involves reproducing pre-determined language through drills, pattern practice, and manipulating structures (Goh & Burns, 2012). This approach revolves around structural accuracy and the application of language forms (Novakova, 2016). The indirect approach concentrates on establishing environments that foster oral interaction and encourage learners to engage in communicative activities, highlighting fluency and authentic language use (Goh & Burns, 2012; Richards, 2008). However, solely relying on either approach can lead to imbalances, with students excelling in one area while underperforming in others (Adem & Berkessa, 2022).

Both direct and indirect methods could be combined under a holistic approach (Thornbury, 2006), comprising three cyclical stages, namely, awareness-raising (where learners identify and familiarise themselves with their knowledge gaps), appropriation (where learners incorporate new knowledge into their existing framework, transitioning from externally-regulated practices to self-regulated ones), and autonomy (where learners apply their newly-acquired knowledge in real-world situations).

With this in mind, integrating debate activities into the L2 classroom could serve as a potentially effective approach to enhancing language development. The debating format inherently promotes interaction and the active use of language (Firmansyah & Vegian, 2019). It provides learners with a platform for practising accuracy and fluency, utilising language in an engaging, real-life communicative context (Noor & Tajik, 2022). The following section provides a critical review of the existing research on the effects of debate pedagogy, examining it from theoretical and practical perspectives.

2.2 Debate as an educational approach: theoretical grounds

Debate is considered a potentially practical educational approach for developing learners' speaking skills (Brown, 2009; Firmansyah & Vegian, 2019; Gatbonton & Segalowitz, 2005). This section explores the definitions of debate, the commonly used procedures and formats in classroom debates, and the theoretical foundations that support debate pedagogy in language development, including the output and interaction hypotheses.

2.2.1 Definition of debate

Debate plays a pivotal role in our daily communications (Brown, 2009). Many important decisions that have far-reaching consequences on both a personal and governmental level are reached following thorough debates (Freeley & Steinberg, 2013). Debate centres around the concept of change, highlighting our innate human tendency to seek improvement and avoid complacency (El Majidi, 2022). It provides a platform for effectively discussing and negotiating these desired changes (Davis et al., 2016).

The definition of debate has evolved over the years (Trapp, 2016). Branham (1991) defined debate as the process of advancing, supporting, disputing, and defending

opinions. This broad definition emphasises the core aspect of debating: supporting one's standpoint with arguments and challenging the opponents' reasoning. Freeley and Steinberg (2013, p. 6) defined debate as 'the process of inquiry and advocacy, a way of arriving at a reasoned judgment on a proposition'. El Majidi (2022) argued that a debate should be equitably organised around a topic of interest, and all sides should be allocated an equal amount of time to articulate their viewpoints and identify the flaws in their opponents' arguments.

Whilst there is no universally agreed definition, the present study summarises debates as the process involving two teams with equal opportunities to present, support, question, and defend arguments for and against a motion, aiming to reach a well-reasoned judgement on a proposition.

Debate differs from ordinary conversation and discussion. Lubetsky et al. (2000) described debating as a highly sophisticated form of immediate and interactive communication that requires an elevated level of discourse skill. This complexity extends beyond ordinary conversation, requiring participants to engage in dynamic and critical listening, and demonstrate advanced language competency and critical thinking (Lieb, 2007). Moreover, Hall (2011) stressed that debates differ from discussions in that they involve taking a prearranged position, either for or against a particular issue. In addition, in contrast to discussions that may lack a clear structure or predefined rules, debates often adhere to structured formats that go through several stages (e.g., constructive speech, rebuttal, and summary) (El Majidi, 2022). Finally, whilst discussions can simultaneously involve the whole class, debates are often undertaken in groups, usually consisting of two to eight debaters.

2.2.2 Development and procedure of debate

In this subsection, an overview of debating as an educational approach, and the frequently used debate procedures and formats are presented.

History of educational debate

The educational debate can be traced back to nearly 2,400 years ago when Protagoras of Abdera, recognised as 'the father of debate', required his students to debate against each other in Athens, Greece (Kennedy, 2009, p. 225). The first English-speaking intercollegiate debate may have been the debate between Oxford and Cambridge university students in the early 1400s (Garrett et al., 1996). The 1800s witnessed a surge in intercollegiate and interscholastic debates, leading to the emergence of tournament debate as a distinguished activity (Darby, 2007). In the early 1900s, the appeal of debate diminished, resulting in a short period of decline (Freeley & Steinberg, 2013). However, the demands for reforms in education sparked a resurgence in interest in debates in the 1980s (Darby, 2007), renewing its reputation as an effective tool for enhancing communication skills and critical thinking.

Debate pedagogy has long been recognised for its value in the L1 context (Tumposky, 2004). In recent decades, the potential benefits of this pedagogy in ESL/EFL learning have been acknowledged, leading to its increasing application in L2 classrooms (El Majidi et al., 2018). In the present study, debate pedagogy in L2 classrooms refers to integrating debates into the curriculum, where students participate in debates on particular topics using ESL/EFL, following predetermined procedures and formats.

Debate procedure and formats

In the process of classroom debates, students should explore pertinent literature on the debate topic, analyse data, formulate solutions or hypotheses before the debate, and deliver their arguments and rebuttals persuasively during the actual debate (Lustigova, 2011). Debaters are expected to follow a certain debate format. Meanwhile, non-debating students, as the audience or judges, are responsible for attentive listening and notetaking, engaging in post-debate discussions, and evaluating the quality of the debaters' arguments (Aclan & Aziz, 2015a). The process of debating thus creates an engaging learning environment that has the potential to enhance students' target language proficiency along with other key skills, such as critical thinking and evaluation.

Many classroom debate formats have been proposed in the literature, and most of them closely resemble or are identical to competitive debate formats (Merrell et al., 2015; Freeley & Steinberg, 2013). Here, some examples frequently used in both L1 and ESL/EFL classrooms are presented. These include the Lincoln-Douglas debate, the public forum format, and parliamentary debates, with debate teams ranging from one-on-one to four-on-four (Brown & Bown, 2014; Trapp, 2016).

The Lincoln-Douglas debate is a one-on-one format (Figure 1), named in honour of the debates between Abraham Lincoln and Stephen A. Douglas during the 1850s (Brown & Bown, 2014). This format comprises four elements: constructive speech (debaters initially define the motion/topic, build the case, and present their arguments), cross-examination (opponents ask each other questions to challenge or clarify the speaker's arguments), rebuttal (debaters respond to and counter the opposing team's arguments), and summary (debaters recap their arguments and highlight their key points) (Freeley & Steinberg, 2013).

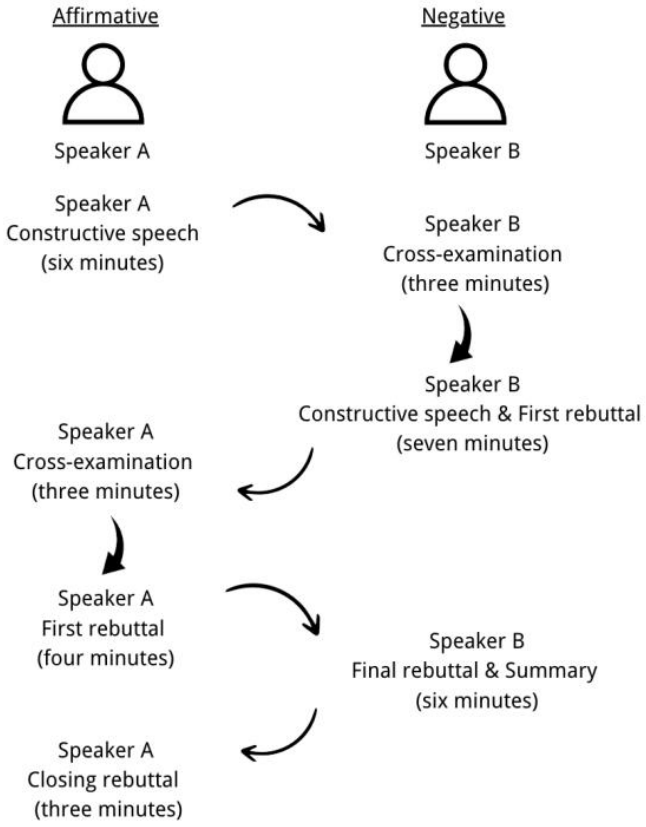

Figure 1. Lincoln-Douglas debate format

(adapted from Freeley & Steinberg, 2013)

The public forum debate is a two-on-two format (Figure 2). It originated in the US during the early 2000s, and is now widely used in secondary schools worldwide (Hannan et al., 2012). Under this format, students deliver constructive speeches and engage in rebuttals. They also participate in 'crossfire' (similar to cross-examinations) with the opportunity to question the opposing team and the 'final focus' where debaters explain why they believe their team should prevail in the debate.

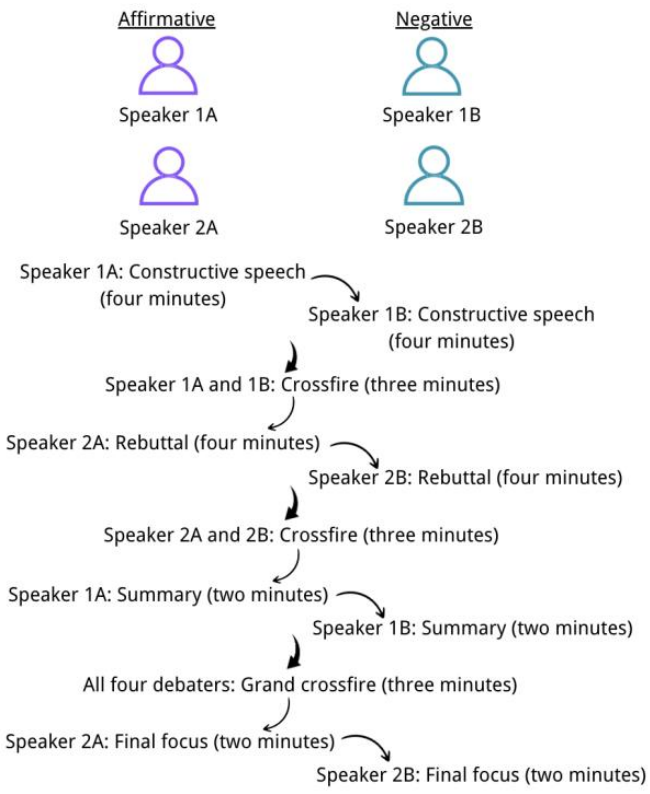

Figure 2. The public forum debate format

(adapted from Freeley & Steinberg, 2013)

Parliamentary debate formats typically refer to the Asian Parliamentary (AP) or British Parliamentary (BP) debates (Syahputra & Chaira, 2020; Eckstein & Bartanen, 2015). These formats are labelled 'parliamentary' as they simulate real-life parliamentary sessions, with debaters taking on roles as government representatives

or members of the opposition (Trapp, 2016). The AP debate is a three-on-three format prevalent in Asian universities (Figure 3).

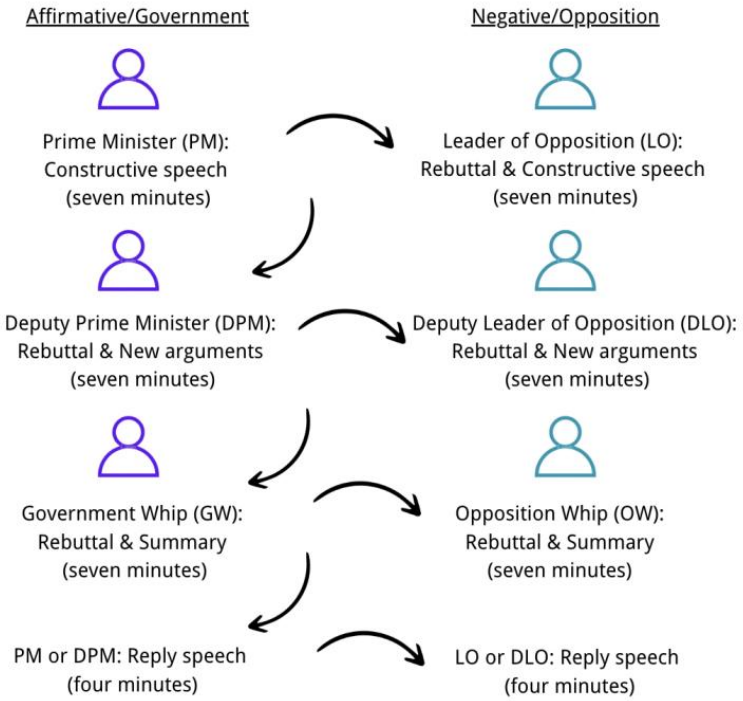

Figure 3. Asian Parliamentary debate format

(adapted from Syahputra & Chaira, 2020)

Meanwhile, the BP debate, also known as 'World Schools Debate', is a four-on-four debate format (Figure 4) (Litan, 2020, p. 37). Originating from British debating societies such as the Oxford Union in the early nineteenth century, the BP debate format is one of the most widely used formats in universities worldwide (Jodoi, 2023). Additionally, parliamentary debates incorporate a feature called Point of Information (POI), which allows debaters from opposing teams to seek clarification or challenge the arguments presented by the speaker currently delivering their speech within a specific timeframe (often between the first and last minute of the speech), and the speaker may either accept the point or decline it (Trapp, 2016).

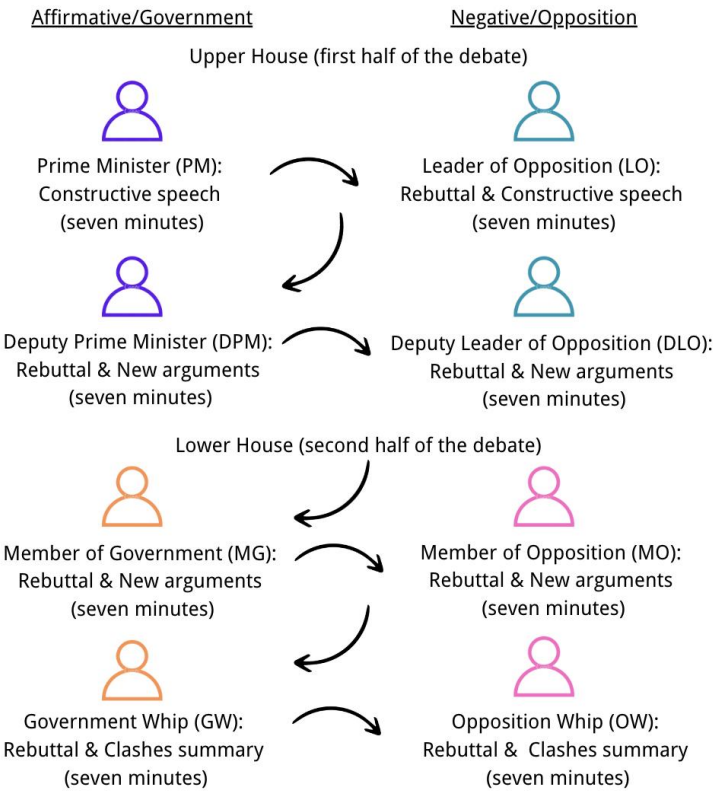

Figure 4. British Parliamentary debate format

(adapted from Trapp, 2016)

As shown from Figure 1 to Figure 4, whilst these formats share certain structural components, including constructive speech, rebuttal, and summary, they also exhibit differences in the number of speakers on each team, the order in which they speak, and the amount of time allocated to each speaker. Following specific rules, these formats are designed to maintain order, efficiency, and impartiality among participants (Freeley & Steinberg, 2013).

However, understanding and implementing these debate formats may be time-consuming and complicated; hence they are considered more appropriate for upper-intermediate and advanced language learners (Knowles, 2023; Brown & Bown, 2014). Nguyen (2020) highlighted this issue from the perspectives of Japanese EFL students, with one student stressed the difficulty of debating intricate topics in English with complicated formats, asserting that it would be challenging even in their native

language. Another student in Nguyen's (2020) study acknowledged the issue associated with providing a rebuttal, which requires a range of skills such as listening, critical thinking, sentence construction, and speaking.

To address these concerns, Lustigova (2011) recommended using simplified debates, which include providing step-by-step guidance prior to the actual debate, selecting suitable topics, and reducing speaking time. These adjustments may facilitate ESL/EFL students with intermediate and lower proficiency levels to participate in debates. Brown and Bown (2014) suggested that teachers should select, modify or design a debate format that best aligns with the needs and comfort level of their class.

2.2.3 Output hypothesis and interaction hypothesis

Numerous theoretical approaches to L2 acquisition provide a rationale for assuming the potential impact of debates on language development, such as the Output Hypothesis (Swain, 2013; Swain & Lapkin, 1995) and the Interaction Hypothesis (Long, 1996, 2018).

Swain and Lapkin's (1995) Output Hypothesis offers a theoretical viewpoint that might explain how it could enhance L2 acquisition when applied to debating. They posited that output gives learners unique opportunities for language processing as it prompts learners to identify their language shortcomings, allowing them to use language forms at the limits of their current capabilities and encouraging conscious reflection on language forms. Participating in two-way output, such as in debates, which includes expressing participants' own arguments and rebutting others, reveals shortcomings in L2 learners' interlanguage systems, thus helping address problematic areas. Benati (2017) asserted that engaging L2 learners in structured collaborative output tasks could facilitate the accurate and appropriate use of language forms and structures. One such task is debating, which naturally generates substantial oral output as participants confront each other's viewpoints, striving to outperform each other's arguments. This process may not only heighten awareness of linguistic

deficiencies but also encourage experimentation with new forms and conscious language use (El Majidi et al., 2018).

Regarding the Interaction Hypothesis, Long (1996, 2018) asserted that interactive tasks are conducive to meaningful negotiations, effectively connecting input and output. Similarly, Ellis and Shintani (2014) posited that interaction provides a rich source of input and output opportunities, promoting internal processing that facilitates acquisition. Interaction creates abundant opportunities for learners to negotiate meaning and form, ultimately contributing to L2 development (Gass & Mackey, 2015; Loewen & Sato, 2018). Debating, as an interactive task, facilitates meaningful interactions at multiple levels. These include the interaction with debated content and between learners and instructors (Brown & Bown, 2014). When involved in debating, students are forced to critically analyse the opponents' reasoning and evidence, aiming to highlight their weaknesses and deficiencies (El Majidi, 2021a). This interactive process requires debaters to search, discern, scrutinise and evaluate arguments, overcome personal prejudices, and identify flaws in the opponents' reasoning lines, helping learners identify and subsequently refine their L2 output to bridge language gaps.

2.3 Debate as an educational approach: pedagogical grounds

Debate pedagogy has been employed in a variety of disciplines, ranging from political science (Omelicheva, 2006), mathematics and statistics (Stewart & Stewart, 2014), and healthcare (Khan et al., 2012; Lampkin et al., 2015; Mamtani et al., 2015), to L2 learning (e.g., Alasmari & Ahmed, 2013; El Majidi, 2022; Zare & Othman, 2015). Moreover, many studies have discussed the benefits that are associated with debate participation, such as critical thinking (Akerman & Neale, 2011; Bellon, 2000; Campo et al., 2023); citizenship (Rutten & Soetaert, 2013; Zorwick & Wade, 2016); disciplinary learning (Anderson, 2016; Goodwin, 2003; Kennedy, 2009); and L2 language proficiency, including reading, listening, writing and speaking (e.g., Brown & Bown, 2014; El Majidi et al., 2018, 2020, 2021a, 2021b, 2023; Jodoi, 2023).

For the purpose of this review, the focus is on studies that have applied debate pedagogy within the ESL/EFL context. The educational outcomes are primarily related to the development of speaking skills, covering both language and affective aspects, as these are often brought up in discussions. Additionally, some typical critiques of this approach are also discussed.

2.3.1 Language aspects of ESL/EFL speaking

Many studies have shown the effects of debate participation in developing ESL/EFL speaking, ranging from overall speaking proficiency to specific aspects of language such as vocabulary, grammar, pronunciation, fluency, accuracy, complexity, and argumentation (e.g., Al-Mahrooqi & Tabakow, 2015; El Majidi, 2022; García-Sánchez, 2020). The reported improvements include subjective perceptions and objective advancements, which are critically reviewed as follows.

Studies have consistently reported that both instructors and students perceived improvements in speaking skills through classroom debate participation (e.g., Aclan & Aziz, 2015a; Lustigova, 2011; Zare & Othman, 2015). For example, Lustigova (2011) reported that teachers observed improvements in students' overall English communicative competence after participating in classroom debates at a Czech Republic university. Likewise, Aclan and Aziz (2015a) interviewed six university students from three Asian countries as a group who had participated in classroom debates for 16 weeks, and their findings indicated that students perceived that their speaking skills, such as vocabulary, grammar, fluency, and accuracy, had improved after the debate intervention.

Whilst these studies provided valuable insights at the tertiary level, El Majidi et al. (2018) conducted a study involving 134 Dutch secondary school students. After a two-year classroom debate intervention, the results from questionnaires, interviews, observations, and speaking performance revealed that both students and teachers perceived improvement in participants' language skills, including listening, reading, writing, and speaking. Amongst these skills, participants perceived that their speaking

skills improved the most, such as the ability to deliver longer speeches and improve fluency (El Majidi et al., 2018). It is important to consider instructors' and learners' perceptions and attitudes in the pedagogical decision-making process (Field, 2017), as positive perceptions and attitudes can influence their motivation to teach and learn, fostering a relaxed and enjoyable learning environment and potentially leading to more efficient L2 acquisition.

Despite showing the positive perceived effects, however, these studies have certain limitations. Lustigova's (2011) findings are based on observations, which could inadvertently incite students' reactivity and observers' expectancy effects. 'Reactivity' means participants might adjust their actions to align with what they believe the researcher wants to see, especially if they are aware they are being observed, whereas 'expectancy effects' refers to the observer unconsciously anticipating the detection of specific behaviours (Cohen et al., 2018, p. 560). These may influence researchers' observations, leading to potential bias. Moreover, findings from Aclan and Aziz (2015a) through a group interview with six students could potentially lead to the 'group think' phenomenon. This means individuals who possess differing opinions might feel discouraged from expressing their views in front of the rest of the group (Cohen et al., 2018, p. 527). Furthermore, all three studies reported perceived effects via self-reported data, which may not necessarily correspond to actual improvements. In addition, these studies lack a control group to compare intervention effects, making it challenging to ascertain the specific effects of debate intervention relative to other educational approaches.

To address these issues, El Majidi et al. (2019a, 2019b) implemented a pre-test-post-test control group quasi-experimental design to investigate the effects of debate intervention on the EFL speaking competence of Dutch secondary school students. The study included five intact classes (n=96) as intervention groups and three classes (n=51) as control groups. The intervention groups engaged in 10 weekly debates, whilst control groups received regular instruction to practice four core language skills - reading, listening, speaking and writing. Students were evaluated via oral opinion

tasks, with different topics randomly assigned to participants before and after the intervention. The collected speech samples were then analysed for speech quantity, fluency, complexity, accuracy, and cohesion (El Majidi et al., 2019a) and the structure and quality of arguments (El Majidi et al., 2019b). The findings revealed that the intervention groups produced more language that was fluent, accurate, coherent, and lexically sophisticated compared to the control groups following the debate intervention. Moreover, the intervention groups outperformed the control groups in oral argumentation skills in terms of argument structure and quality. They concluded that debate pedagogy has a facilitative impact on improving L2 speaking proficiency.

These recent studies provide more robust empirical evidence to support the effectiveness of debate pedagogy in L2 speaking development. However, the intervention groups' superior speaking outcomes may also be attributed to the practice effect. This means that the intervention groups, having had more opportunities to practice speaking, could potentially be advantaged by the speaking tests used to assess their oral performance. Hence, it is challenging to determine whether the observed improvements are the product of debate intervention itself or affected by other variables, such as the results of increased practice.

2.3.2 Affective aspects of ESL/EFL speaking

In addition to the language perspective, research has indicated that debate pedagogy can potentially influence the affective aspects of L2 speaking, such as increasing self-confidence and reducing anxiety (e.g., Jodoi, 2023; Jouini, 2019; Purnama, 2018).

Purnama (2018) and Jouini (2019) reported that over 90 percent of EFL participants felt more confident speaking English after participating in classroom debates. Debates help bolster students' self-confidence by compelling them to articulate and defend their position against the opposition under pressure (Freeley & Steinberg, 2013).

Additionally, participants in Aclan and Aziz (2015a) who engaged in 16 weeks of classroom debates reported that they initially experienced anxiety, but this feeling gradually subsided as their frequency of participation in debates increased. Similarly, Khalil (2018) also discovered that students' L2 speaking anxiety reduced and speaking performance improved after participating in eight debating sessions.

However, the above studies are mainly based on self-reported data, which might lead to social desirability bias. This means that participants have a tendency to provide responses they believe are more socially desirable or acceptable rather than responses that reflect their true feelings or opinions (Grimm, 2010). Hence, caution should be exercised when interpreting these findings.

Collectively, the potential benefits of debate pedagogy in L2 speaking in both linguistics and affective aspects are apparent, yet reaching a definitive conclusion about its direct impacts on ESL/EFL speaking development requires careful consideration and further robust research.

2.3.3 Critiques of debate pedagogy

Whilst the previous sections discussed the positive aspects of debate pedagogy, some scholars have also raised concerns about this educational approach (e.g., Pica, 1996; Tannen, 1998; Tumposky, 2004). These mainly include the claims that debate encourages dualism, focuses excessively on competition, and fails to foster prosperous negotiations (El Majidi, 2022).

Debates have been critiqued for promoting dualism. For example, Tannen (1998) stated that debates tend to emphasise opponents' logical weaknesses and manipulate facts to gain an advantage. Tumposky (2004) also suggested that debates bolster a dualistic bias and minimise multiple perspectives, often trivialising complex ideas and favouring ones that support a particular standpoint. However, Jackson (1973) and Scannapieco (1997) suggested that debates inherently expose students to situations demanding tolerance of differing viewpoints. MacArthur et al. (2002) also argued that

debates are a promising avenue for students to explore diverse perspectives on contentious issues. Kennedy (2007, 2009) reported that debates aided students in developing empathy and altering their views, with a notable percentage of participants experiencing viewpoint changes. As such, debates could encourage tolerance, empathy, and perspective shifts rather than dualistic viewpoints.

The competitive nature of debates was another target of criticism, with some arguing that it creates a confrontational environment that may be uncomfortable for some students (Tannen, 1998). Moreover, students might oversimplify complex topics in the pursuit of debate victory (Tumposky, 2004). However, Goodwin (2003), for example, evaluated participants' views on the competitive nature of debates, finding that only a small minority (two out of 52) expressed unease due to the competitive aspects. Crowell and Kuhn (2014) argued that whilst the initial focus in debates may be on winning, this emphasis typically shifts over time towards the quality of arguments and rebuttals. In line with this, Merrell et al. (2017) advised instructors against awarding extra points for winning a debate to reinforce the idea that the goal is not to defeat an opponent but to demonstrate the capacity to construct, apply, and defend well-substantiated arguments.

Regarding negotiation, Pica (1996) posited that reflective activities such as debates do not foster substantial negotiation, primarily because more assertive learners tend to dominate the discussions, often marginalising the contributions of their less assertive peers. Conversely, studies by Stewart and Pleisch (1998) and Dundes (2001) indicated that debates can inspire even the less assertive students to participate in the interaction actively. As El Majidi (2022) argued, this can be achieved via well-structured debates that ensure equitable time distribution to each participant to stimulate productive negotiations. The reflective nature of debates can propel these negotiations to further depths as participants challenge each other's viewpoints, particularly during rebuttals.

Collectively, the use of debate pedagogy has been both praised and criticised. Whilst the criticisms raised are credible, they did not provide conclusive evidence to prove that debate pedagogy is ineffective. Instead, the contrasting views towards this educational approach highlight the need for meticulous research design and implementation to ensure that debates promote constructive engagement with diverse perspectives rather than mere opposite views and competitions.

2.4 Gaps in the literature

The importance and challenges of teaching and learning ESL/EFL speaking were explored in Section 2.1. Additionally, the potential effects of debate pedagogy in developing ESL/EFL speaking skills were critically reviewed from theoretical and empirical perspectives in Sections 2.2 and 2.3. Based on these points, the prior studies still reveal some research gaps.

Despite many studies highlighting the benefits of debate pedagogy in enhancing ESL/EFL speaking skills, both in terms of linguistic competency and affective dimensions, the existing literature lacks a comprehensive and systematic examination of several crucial aspects. These issues include an exploration of how widely the debate pedagogy has been employed across different geographical and educational contexts; the types of research design adopted; the implementation of different types of debate pedagogy; the measurement instrument and speaking outcomes under debate pedagogy; its effectiveness in ESL/EFL speaking development via a thorough evaluation of the quality and trustworthiness of the relevant empirical evidence; and its applicability to broader contexts.

These issues indicate that a more rigorous and exhaustive investigation is required to provide an in-depth, reliable, and nuanced examination of the effectiveness of debate pedagogy in terms of enhancing ESL/EFL speaking skills. A systematic review could be a feasible method to achieve this examination, the rationale for which is elaborated in the following chapter.

As far as the researcher is aware, no systematic review has yet addressed the aforementioned issues. Whilst there are two existing systematic reviews evaluating the effectiveness of debate pedagogy (Ang et al., 2019; Carinanos-Ayala et al., 2021), they are primarily contextualised among university health professional students, and it remains unclear whether these students were ESL/EFL learners. Moreover, Ang et al. (2019) and Carinanos-Ayala et al. (2021) did not specifically scrutinise ESL/EFL speaking skills, instead encompassing a broad range of general competencies, including communication skills, content understanding, critical thinking, research skills, and teamwork.

Therefore, this systematic review seeks to supplement the research on debate pedagogy in ESL/EFL speaking development. This will be accomplished by evaluating the scope of research carried out across various settings, synthesising and critically appraising the findings of the relevant studies, and consequently informing effective teaching practices and future research in this field.

2.5 Research questions

To address the issues in the literature presented in the previous section, this systematic review proposes the following research questions:

RQ1: What is the extent of empirical research investigating debate pedagogy in the context of ESL/EFL speaking skills?

RQ2: What are the effects of debate pedagogy on ESL/EFL speaking skills?

Chapter 3. Methods

This study adopted a systematic review approach to address the research questions and executed an exhaustive exploration of the relevant literature. Whilst traditional literature reviews can provide valuable insights, they are sometimes unclear in terms of how authors identify potentially eligible literature and how they decide which literature to include and which to exclude, which could lead to biased reporting (Haddaway et al., 2015). Additionally, traditional reviews may not have a systematic or formalised approach to appraise the included studies' methodological quality critically.

In comparison, a systematic review aims to gather as much relevant research as possible to address specific research questions (Boland et al., 2017). According to the Evidence for Policy and Practice Information Centre (EPPI Center, n.d.), systematic reviews: adhere to a set of explicit and transparent methods; comply with a standardised set of research stages; are replicable, accountable, and updatable; and systematically evaluate the quality of primary evidence, aiming for comprehensive and trustworthy results, and minimise biased reporting. It is imperative to understand the significance of this systematic approach as these reviews amalgamate research findings into an easily accessible format for those who need to make policy or practice decisions (Zawacki-Richter et al., 2020).

The book by Boland et al. (2017), entitled *Doing a Systematic Review: A Student's Guide* was used as a reference manual for this project since it caters specifically to the needs of graduate students. Additionally, *Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols* (PRISMA-P), developed by Shamseer et al. (2015), was employed to formulate an effective protocol for guiding the research process, whilst the report was written up in line with PRISMA 2020 reporting guidelines (Page et al., 2021).

Before undertaking this review, I searched various academic systematic review databases to determine whether the same topic had been addressed in a previous

systematic review. These include the International Database of Education Systematic Reviews (IDESR, n.d.), the EPPI Centre Database of Educational Research (EPPI-Centre, n.d.), the Campbell Collaboration Online Library (The Campbell Collaboration, n.d.), and the International Prospective Register of Systematic Review database (PROSPERO) (The University of York, n.d.).

The search was conducted on these databases using the terms ‘debate’ and ‘speaking’, yielding 126 reviews that were either published or ongoing. However, none of the prior reviews aligned with the objective of the present study. Of the 126 reviews, 85 were unrelated to debate pedagogy, and 41 were unrelated to EFL/ESL speaking development. There were systematic reviews examining the effects of other interventions on L2 speaking skills, such as digital storytelling (Viknesh et al., 2021), drama-based activities (Faitaki et al., 2023, published protocol), and flipped learning approach (Santhanasamy & Yunus, 2022), but none of them investigated debate pedagogy. Additionally, there were two systematic reviews investigating the effectiveness of debate pedagogy (Ang et al., 2019; Carinanos-Ayala et al., 2021), but the participants were university students in health professions, and they did not specify whether they were EFL/ESL learners, whilst the outcomes did not focus on EFL/ESL speaking skills but rather a wide range of capabilities, such as L1 communication skills, content learning, critical thinking, research abilities, and teamwork. Therefore, it can be inferred that this systematic review investigating the effectiveness of debate pedagogy on enhancing EFL/ESL speaking skills contributes novel insights to this field not covered by prior works.

To prevent duplication of effort and ensure the systematic reviews’ transparency and reproducibility, this review’s protocol was registered with IDESR on March 1st, 2023. The entire protocol can be accessed online at https://idesr.org/article/IDESR000065 (Appendix A).

3.1 Eligibility criteria

Eligibility for inclusion in this review was determined by a set of inclusion and exclusion criteria. The parameters that determined inclusion or exclusion included bibliographic information, publication date, publication status, language, study design, setting, population, intervention, and outcome. A comprehensive list of the inclusion and exclusion criteria, and their underlying justification, can be found in Table 2.

Table 2. Eligibility criteria

Item	Inclusion and exclusion criteria	Rationale
Bibliographic information	Include studies with a complete reference.	Without a complete reference, retrieval of works is infeasible.
	Exclude studies with an incomplete reference.	
Date of publication	Include studies published on any date.	To collect the full scope of research available.
	Do not exclude studies based on their publication date.	
Publication status	Include peer-reviewed articles, organisation reports, master's and doctoral theses, and conference proceedings.	To minimise the impact of publication bias on this topic, this review includes both peer-reviewed and grey literature (e.g., conference proceedings, theses, policy statements).
	Do not exclude studies based on their publication status.	

Language	Include studies published in any language. Do not exclude studies based on their written language. However, if the main text of the eligible study is written in languages other than those the reviewer can read (i.e., English and Chinese), the reviewer would seek to obtain a translation of the text or find someone who can read this language to assist with data extraction.	As the review focuses on ESL/EFL learners, it is possible that eligible studies were written in non-English languages; thus, limiting the review to studies published only in English may lead to a systematic neglect of a particular body of research.
Study design	Include all types of study design. Do not exclude studies based on research design, but exclude secondary research such as systematic reviews, meta-analyses, or commentaries.	Considering the dearth of research in this field, eliminating one kind of study design poses the risk of systematically ignoring a body of work and restricting the scope of this review.
Setting	Include typical classroom settings (e.g., primary and secondary schools, and universities), where debate is used as a pedagogy in EFL/ESL.	This review focuses on classroom settings where debate is used as a pedagogical tool. Findings from studies conducted in other settings, such as debate societies or debate competitions

	Exclude other settings, such as debate societies, debate competitions or tournaments, and out-of-the-classroom/school debates.	or tournaments, may have different purposes and may not generalise to typical classroom settings.
Population	Include typically developing EFL/ESL learners from primary school to graduates with no age restrictions.	This review focuses on typically developing ESL/EFL population, thus findings from non-EFL/ESL speakers and non-typically developing learners will not be included in this review.
	Exclude non-EFL/ESL learners and non-typically developing learners, contexts in which all learners/participants are non-typical development (e.g., special educational needs, developmental language disorder).	
Intervention	Include studies using classroom debate as an intervention.	This review focuses on the effects of using classroom debates as a pedagogical tool; thus, the intervention should include classroom debate.
	Exclude studies with no particular emphasis on classroom debates.	
Outcome	Include studies that report quantitative (e.g., pre-post test scores) and/or qualitative data (e.g., attitude, perceptions, and	Concrete data is essential for a synthesis of findings in this field of research.

	experiences). Include outcome measures related to speaking proficiency in the target language, such as speech quantity (e.g., number of words, speaking time), fluency, accuracy, lexical complexity, syntactic complexity, pronunciation, argumentation, confidence, anxiety, and motivation.	
	Exclude studies that only describe debate pedagogy or are thought pieces/theoretical discussions about the potential benefits of debate pedagogy – i.e., that do not report empirical outcome data on the effects of debate pedagogy - (qualitative or quantitative).	

3.2 Information sources

The databases used for searching in this systematic review can be found in Table 3, which includes databases spanning education, applied linguistics, psychology, and multidisciplinary sources. The searches were also undertaken for grey literature such as conference proceedings, government documents, and master's and doctoral theses to reduce the influence of publication bias, namely, where studies finding significant effects are preferred in peer-reviewed publications and reports with less profound results are less likely to be submitted and accepted for publication. Due to time

constraints, the review primarily relied on electronic database searches, and manual searching was not conducted. All databases listed below (Table 3) were accessible electronically with a subscription from the University of Oxford's Bodleian Library.

Table 3. List of databases

Discipline	Database
Education	British Education Index, Australian Education Index
Psychology	PsycINFO
Multidisciplinary	ProQuest Social Science Premium Collection, Scopus, Web of Science Core Collection
Grey literature	ProQuest Dissertation & Theses Global, EThOS, OpenGrey

The preliminary search was supplemented with forward and backward citation searches. Backward citation here means scrutinising the reference section of any studies that fulfilled all inclusion criteria, to identify other potentially eligible studies. For the forward citation search, databases such as Scopus, Web of Science, and ProQuest Social Science Premium Collection were used as they provide the 'cited by' function, which allows the researcher to identify subsequent research that has cited a previously identified eligible study, and then to screen them for eligibility for this review. I also searched on Google Scholar to discover additional eligible studies not returned in the aforementioned databases. Finally, automatic updates on searched databases were implemented to incorporate up-to-date published articles during research period. The main search was last updated on May 20, 2023.

3.3 Search strategy

A well-designed systematic review should strike a balance between sensitivity and specificity. The scope of the search should be broad enough to avoid missing eligible studies while simultaneously being narrow enough so as not to necessitate sifting through excessive irrelevant research (Boutron et al., 2017). The search terms are, therefore, crucial in establishing this balance. Various search strategies were experimented with to ensure reliability while achieving optimal outcomes, including using the thesaurus functions of the databases used and consulting an experienced research librarian in the Department of Education at the University of Oxford to formulate the initial search. The strategy for searching was honed by appraising the titles, abstracts and keywords retrieved and including or excluding different search terms, such as plausible synonyms or associated terms.

Database search options were chosen, where available, to limit the search to abstracts, keywords and titles. To account for the terminology used in various educational contexts, a Boolean search string was constructed by combining different terms in each category using the 'OR' operator and connecting different fields with 'AND'. The final Boolean search string is detailed in Table 4. The pilot search correctly identified two previously identified candidate articles (El Majidi et al., 2021a, 2021b) that met all inclusion criteria, thereby validating the appropriateness of the search terms.

Table 4. List of search terms

Participants	EFL OR ESL OR EAL OR ELL OR ELT OR English* OR L2 OR "second language*" OR FL OR "foreign language*"
AND	
Intervention	debat*
AND	

Outcome	speak* OR spoke* OR communicat* OR oral* OR speech* OR discuss* OR express* OR argument* OR fluency OR accuracy OR complexity OR lexical OR syntactic OR pronunciation OR confidence OR anxiety OR motivation

3.4 Data management

Prior to the review process, I collaborated with my supervisor and a PhD candidate in the Department of Education at the University of Oxford on the software (Rayyan, ASReview and Zetero) that can be useful for the systematic review to ensure familiarity with the screening functions and process. After completing the final searches on May 20, 2023, bibliographic information and abstracts were retrieved as RIS files and uploaded to Rayyan, a program for managing systematic reviews (Ouzzani et al., 2016), to resolve duplication issues.

After removing duplicate records, I started screening the titles and abstracts. The Active learning for Systematic Reviews (ASReview), an open-source machine learning-aided pipeline for systematic reviews (van de Schoot et al., 2020; van Haastrecht et al., 2021), was used to accelerate the process of title and abstract screening. The titles and abstracts were reviewed against the eligibility criteria and tagged for full-text review if they did not expressly breach one or more of the exclusion criteria.

Following the initial screening, the full texts of reports that could not be excluded on the basis of the information contained in their titles and abstracts were downloaded and organised in the Zotero, a reference manager. After full text screening, the data for studies that met all inclusion criteria were extracted and recorded in a Microsoft Excel spreadsheet. All evaluations were performed on a MacBook Pro running Mac OS Monterey.

3.5 Selection process

This section describes the selection process, including title and abstract screening, full-text screening, and how I ensured screening reliability.

3.5.1 Title and abstract screening

ASReview was used to screen titles and abstracts. The choice was made because this software offers advantages over other review tools such as Rayyan, particularly when handling a large initial set of results. In contrast to manual screening, which can be time-consuming, ASReview employs active learning, namely, a learning algorithm that selects the most pertinent data points for analysis (ASReview LAB developers, 2023). This approach allows researchers to efficiently identify the most relevant records while maintaining transparency and ensuring high quality throughout the process (van de Schoot et al., 2021). Details about how the ASReview was operationalised and how it was applied in this systematic review to generate preliminary screening results can be found in Appendix B.

3.5.2 Full-text screening

After the initial screening, 76 studies were identified using ASReview for full-text screening. For studies that were not accessible online or through the Bodleian Library, I contacted the study's primary authors to request a copy. Only studies that explicitly satisfied all inclusion criteria were chosen for the data extraction. If a given study breached a criterion, the screening process ended for that report, and the corresponding criterion was recorded. The results of the search and screening process can be found in Chapter 4.

3.5.3 Ensuring screening reliability

To ensure screening reliability, a second reviewer familiar with applied linguistics and debate pedagogy was invited to conduct a dual-reviewer blind screening. Due to time constraints, the second reviewer could not conduct all abstract and full-text

screening. Thus, she screened 10 percent of the titles and abstracts. The second reviewer was informed of the purpose of, and criteria for, inclusion and exclusion in the present study. During the screening process, both reviewers were blind to the judgements of the other. The two reviewers met to discuss the discrepancies and reach a consensus on addressing them.

A Kappa score is used to measure the inter-rater agreement between two reviewers during the screening process. In educational studies, a Kappa score ranging from 0.61 to 0.80 is the generally accepted threshold for agreement (Frey, 2018). The initial screening process exhibited an inter-rater reliability of 0.75, suggesting that the eligibility criteria applied were reasonable and that the level of agreement between the two reviewers in their judgements was acceptable.

Moreover, I also conducted 10 percent title and abstract screening in Rayyan to compare the results with those obtained from ASReview. It was found that there were 21 potentially relevant studies identified through screening 10 percent of titles and abstracts in Rayyan, 20 of them were included in the 76 relevant studies retrieved from ASReview. These steps provide evidence of the reliability of using ASReview in this review.

3.6 Data extraction

After completing the initial search, a data extraction form was designed expressly for this review (see Appendix C for the detailed data extraction items). The Cochrane data extraction form was used as the model since it is adaptable and suitable for various study designs (Cochrane Effective Practice and Organisation of Care, 2017). The data extracted include the essential elements of participants, intervention, comparison, and outcomes (Petticrew & Roberts, 2006) relevant to this review. Data were extracted and recorded in a Microsoft Word document when a study fulfilled all the inclusion criteria. While it is generally agreed that having multiple reviewers extract data as an internal check and balance when preparing a systematic review is

desirable, a shortage of time impeded the possibility of having another reviewer perform data extraction despite being included in the protocol.

3.7 Trustworthiness appraisal

Primary studies may suffer biases from various sources, such as the conduct of primary study authors, conflicting interests, designs employed for the studies, and actions taken by reviewers (Boutron et al., 2019). Therefore, it is imperative to evaluate the trustworthiness associated with each study in a systematic review as it can provide valuable information regarding the credibility of the overall findings.

The trustworthiness appraisal is a critical aspect that requires tools capable of accommodating the research design's structure (Dixon-Woods et al., 2005). Considering the various research designs employed in the studies found in this review, an evaluation of each study's trustworthiness was carried out using the Mixed Methods Appraisal Tool (MMAT) (Hong et al., 2018). The MMAT is found to be particularly suitable for evaluating studies with diverse methodologies within systematic reviews (Crowe & Sheppard, 2011).

Using the MMAT (see Appendix D), methodological quality can be assessed in five design categories: qualitative, randomised controlled trials (RCTs), non-randomised studies, quantitative descriptive studies, and mixed methods studies (Hong et al., 2019). If a study adopts the mixed methods strategy, multiple categories will be assessed. Across the five design categories, the MMAT is used to score reports of research against five design-specific quality criteria as 'yes,' 'no,' or 'can't tell,' as well as a comments box to explain why a rating was awarded.

When reporting MMAT results, Hong et al. (2020), the creators of MMAT, offer the following scoring recommendation (see Appendix E). There are five criteria for each category of study design, and the overall quality score for an eligible study can be calculated using descriptors such as stars (*):

5***** when all five quality criteria are met.

4**** when four of five criteria are met.

3*** when three criteria are met.

2** when two criteria are met.

1* when one criterion is met.

For mixed methods studies, there are 15 criteria to evaluate instead of five, the overall quality is determined by the lowest score among the study components:

1* when the score for the qualitative component (QUAL), quantitative component (QUAN), or mixed methods component (MM) is 1.

2** when the score for QUAL, QUAN, or MM is 2.

3*** when the score for QUAL, QUAN, or MM is 3.

4**** when the score for QUAL, QUAN, and MM are all 4.

5***** when the score for QUAL, QUAN, or MM is 5.

Hong et al. (2020) argued that it is important to be transparent and clearly describe how the results of the appraisal were interpreted and used in the review. Given that critical appraisal involve making judgements, it is recommended to have at least two reviewers independently participate in the appraisal process (Hong et al., 2018). To uphold the reliability of the assessment and due to time limitations, a second reviewer was involved to scrutinise four representative studies. These studies included one qualitative study, one randomised controlled trial, one quantitative non-randomised study, and one mixed methods study. This decision was made because the studies included in the review employed these four types of study design.

As with title and abstract screening, a Kappa score of 0.61 to 0.80 for the agreement was required. The inter-rater reliability for each category was 0.7, indicating that there was an acceptable level of consistency between different reviewers when

assessing the same set of studies, thus ensuring the reliability of the evaluation process.

3.8 Data synthesis

Considering the significant variability in interventions and outcomes across studies, it would be inappropriate to employ meta-analysis in this review (Petticrew & Roberts, 2006). Instead, I adopted narrative synthesis to evaluate the quality and findings of the studies, given the heterogeneous nature of outcomes within this specific group and the limited number of studies that report comparable measures. In the following chapter, the present study follows the general framework for the narrative synthesis of heterogeneous literature suggested by Popay et al. (2006), which includes four main elements:

(1) Develop a theory of how the intervention works, why and for whom.

(2) Develop a preliminary synthesis of the findings of included studies.

(3) Explore relationships in the data.

(4) Assess the robustness of the synthesis.

Chapter 4. Results

The chapter is structured as follows. Section 4.1 presents the results of the search and selection process. Section 4.2 describes the general characteristics of included studies. Section 4.3 assesses the trustworthiness of individual studies in terms of their methodological quality before presenting the findings in Section 4.4.

The first two sections aim to provide preliminary answers to the first review question regarding the extent of empirical research investigating debate pedagogy in the ESL/EFL speaking skills context, whilst the following two sections address the second review question concerning its effects.

4.1 Results of the search and selection process

The PRISMA flow diagram (Figure 5) outlines the search, screening and selection process that resulted in the 29 eligible studies for the review.

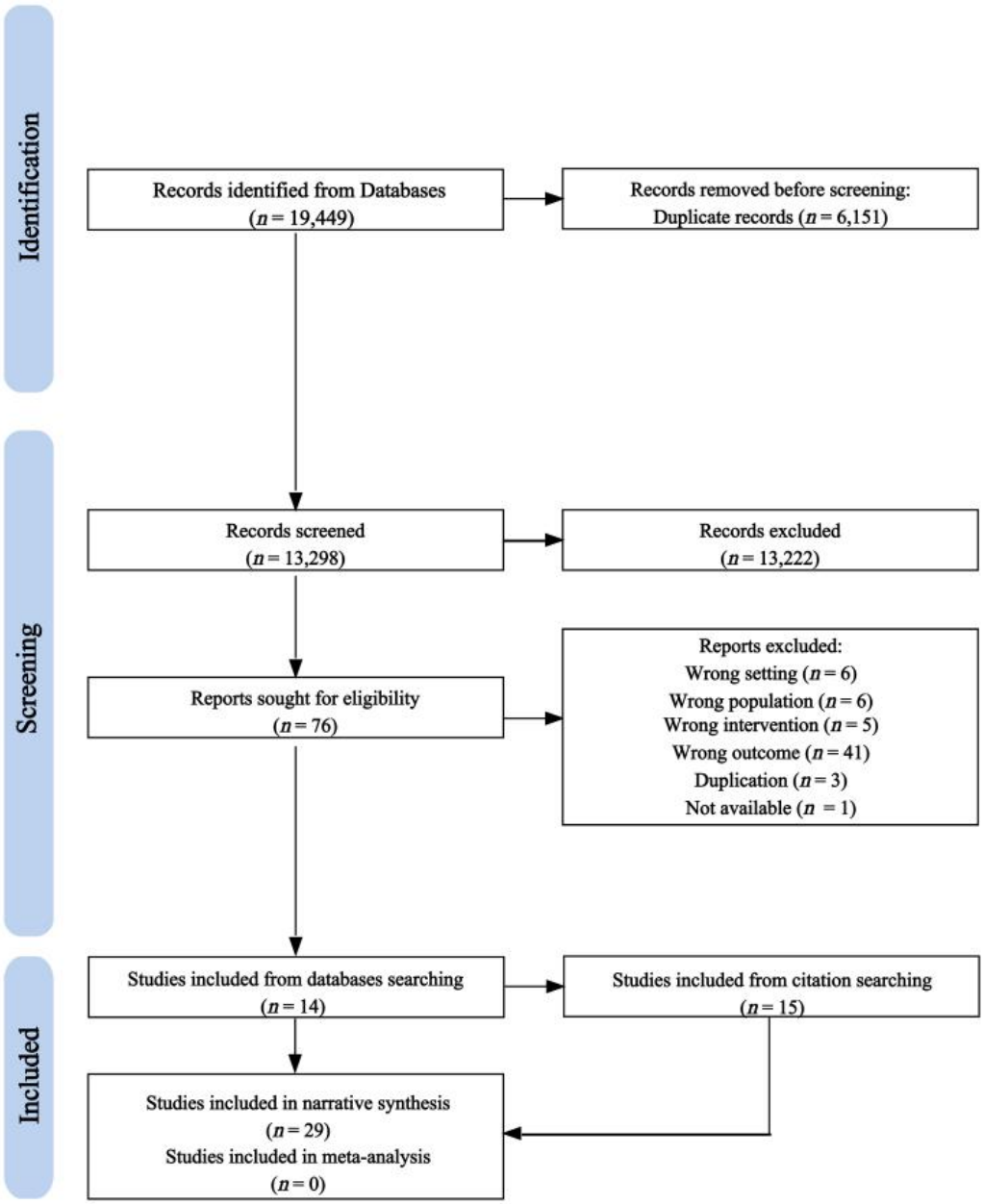

Figure 5. Flow diagram of the screening process

(Template from Moher et al., 2009)

The search was conducted through nine databases (see Section 3.2), which returned 19,449 initial results, before the number was reduced to 13,298 after removing duplicates. Following the initial screening, 76 studies were identified from ASReview for full-text screening. Among these 76 studies, 14 were included in this review. Meanwhile, 62 studies were excluded for the following reasons: wrong setting (n=6),

wrong population (n=6), wrong intervention (n=5), wrong outcome (n=41), duplication (n=3), or not available (n=1). One potentially relevant doctoral thesis by Wilson (2022) could not be downloaded for full-text screening due to publication security issues after confirmation with the original author. I will explain the limitations imposed by the unavailability of this study and evaluate how its absence may have affected the overall findings of the review in Section 5.4.1, relying on the information provided in the abstracts.

Reference lists in the 14 included studies were screened for additional relevant research, and subsequent citations of eligible studies were reviewed for potential inclusion. This process led to an additional 15 studies from backward and forward citations. As a result, the total number of studies included in the review was 29. To facilitate the subsequent description, each included study was assigned an ID number following the alphabetic ordering of authors (Table 5). The completed references of the 29 studies are detailed in Appendix F.

Table 5. Study ID number

ID	Study
1	Aclan & Aziz (2015a)
2	Al-Tamimi (2017)
3	Bahrani et al. (2020)
4	Baso (2016)
5	Brown (2012)
6	Dehham (2021)
7	Desita et al. (2017)
8	El Majidi et al. (2018)

9	El Majidi et al. (2021a)
10	El Majidi et al. (2021b)
11	Elmiyati (2019)
12	Fauzan (2016)
13	García-Sánchez (2020)
14	García-Sánchez & Gimeno-Sanz (2022)
15	He (2015)
16	Iman (2017)
17	Kassem (2021)
18	Khalil (2018)
19	Lustigova (2011)
20	Nobuhara (2016)
21	Noor & Tajik (2022)
22	Putri & Rodliyah (2021)
23	Rosas-Maldonado et al. (2020)
24	Sabbah (2015)
25	Shamsudin et al. (2017)
26	Shan (2005)
27	Tianame et al. (2019)
28	Yulia & Aprilita (2017)

29	Zare & Othman (2015)

4.2 General characteristics

This section presents an overview of the general characteristics of the 29 included studies, including their publication details, geographical and educational contexts, research designs, sample size, interventions, methods of data collection and reported outcomes. Appendix G provides a supplementary table with detailed information on these studies. Their characteristics are summarised in the following sub-sections.

4.2.1 Publication details

Publication year

Figure 6 illustrates the trend over time using five-year bands from 1999 to 2023. The graph shows that debate pedagogy in L2 speaking is a relatively new but rapidly growing research field as time progresses. The superscript numbers below represent the study ID, unless denoted otherwise.

The earliest eligible study[26] was published in 2005, followed by two studies[5,19] published between 2009 and 2013. A surge in academic research interest occurred from 2014 onwards, where the number of studies increased to 14 between 2014 and 2018, followed by 12 studies between 2019 and 2023. The most recent eligible study[21] was published in December 2022.

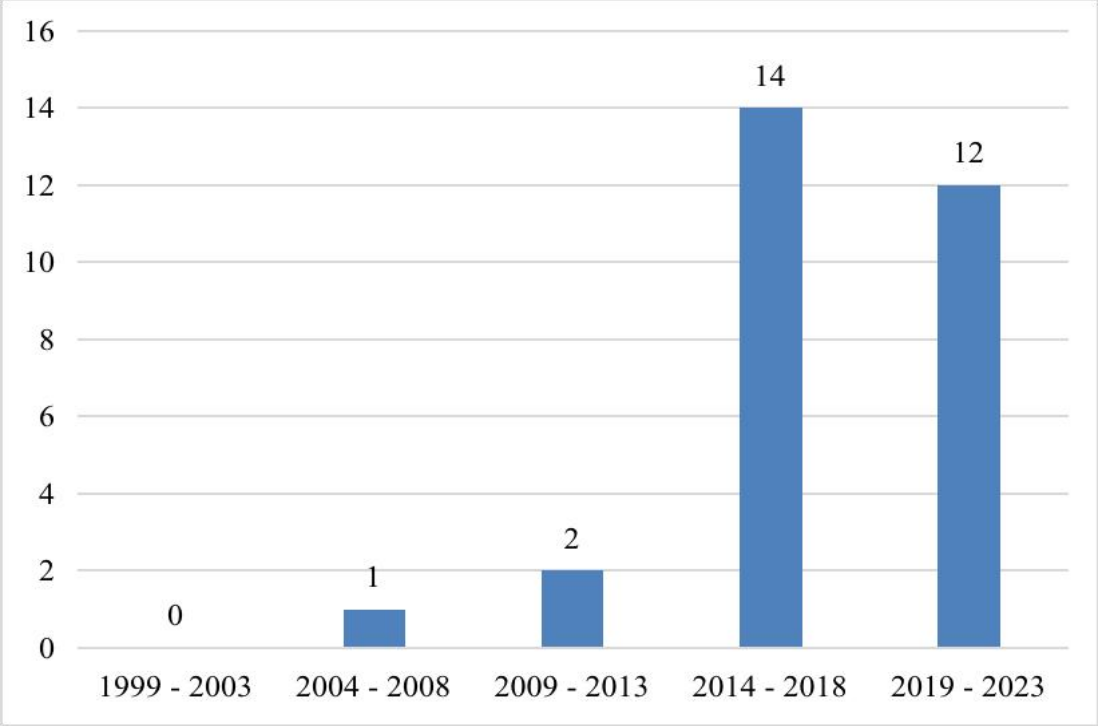

Figure 6. Publication year

Publication status

Figure 7 shows that most of the included studies were journal articles (n=23), followed by four master's theses and two conference proceedings.

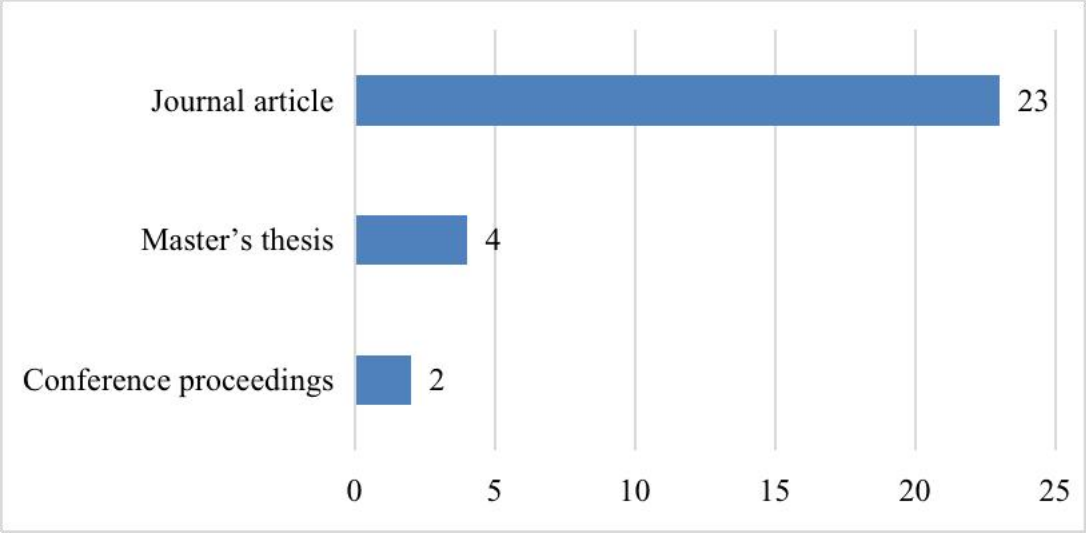

Figure 7. Publication type

4.2.2 Geographic and educational contexts

The research on debate pedagogy in L2 speaking was conducted in various contexts worldwide. Table 6 below categorises the location of studies into Asia, Europe, the Middle East, South America and cross-continental study. The number in parentheses indicates the quantity of studies conducted in that region.

Two studies were conducted across countries, with one study[1] gathering data across the Association of Southeast Asian Nations (ASEAN), including Indonesia, Malaysia, Myanmar, the Philippines, and Thailand, and the other[5] in Russia and the US. The rest were conducted at the single-country level. Asia has the largest body of studies, with Indonesia being the most widely represented country (n=9). Conversely, no study was found in Africa.

Regarding the education phase, the majority of studies (n=18) focused on participants at the tertiary level, whilst the rest were at secondary schools (n=11). No eligible studies were found at primary schools.

Table 6. Location of studies and education phase

Region	**Asia**	**Europe**	**The Middle East**	**South America**	**Cross continental**
	Afghanistan (1)	The Czech Republic (1)	Iraq (1)	Chile (1)	Russia and the US (1)
	ASEAN (1)	Spain (2)	Palestine (1)		
	China (2)	The Netherlands (3)	Saudi Arabia (1)		

	Indonesia (9)		Turkey (1)			
	Japan (1)		Yemen (1)			
	Malaysia (2)					
Total	**16**	**6**	**5**	**1**	**1**	29
Education phase						
Secondary	6	3	2	/	/	**11**
Tertiary	10	3	3	1	1	**18**
Total	16	6	5	1	1	**29**

4.2.3 Research design

The research design of included studies are highly heterogeneous. They were categorised based on the original authors' description and Cohen et al.'s (2018) classification of research methodology.

Figure 8 shows that approximately 38 percent (n=11) employed the one-group pre-test-post-test pre-experimental design, often conducted in an intact class (i.e. an entire class was assigned to specific treatments) and without a control group for comparison. Nearly 28 percent (n=8) of studies adopted the pre-test-post-test quasi-experimental design, which has a control group but lacks random allocation of participants into experimental and control groups. This was followed by the case study, with 14 percent (n=4) of studies collecting quantitative and qualitative data and 10 percent (n=3) only collecting qualitative data. Three studies employed the pre-test-post-test control and experimental group design, namely RCTs, which randomly allocated participants into experimental and control groups. In addition, of 29 studies, three

only collected qualitative data, 15 reported quantitative data, and 11 included both quantitative and qualitative data.

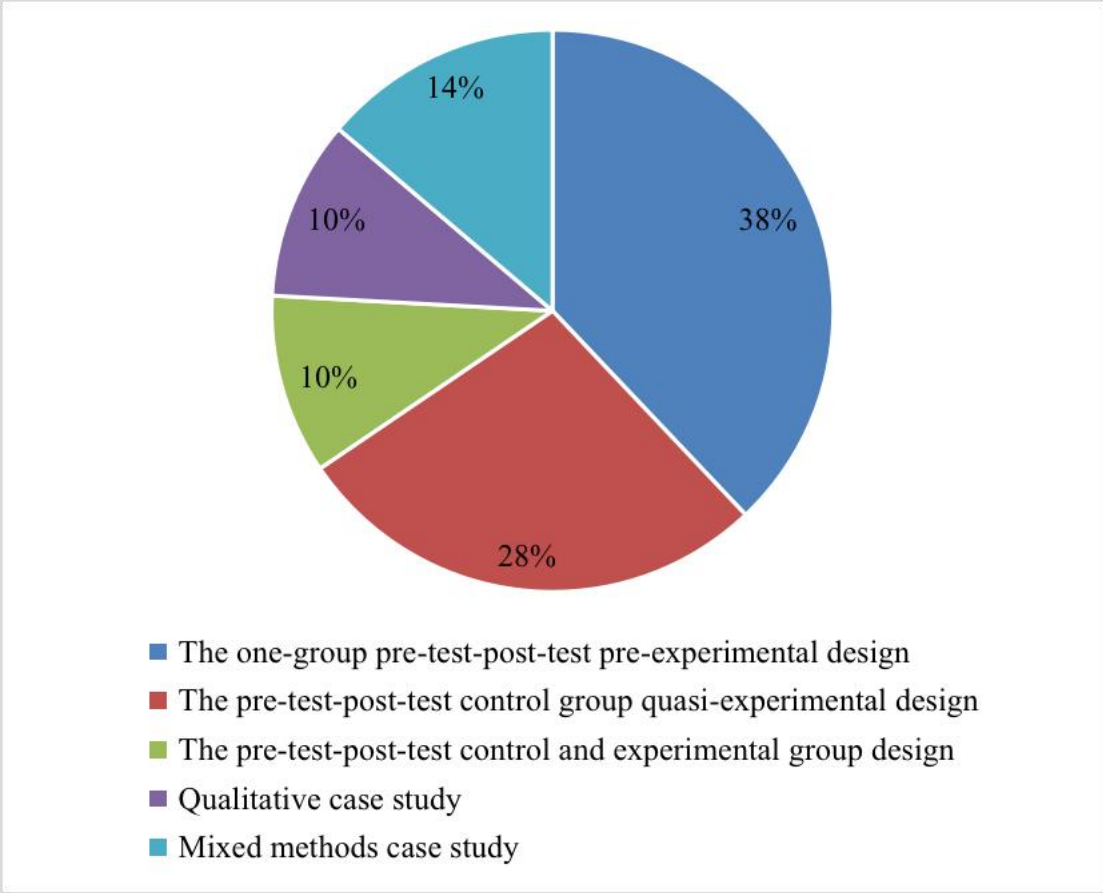

Figure 8. Research methodology

4.2.4 Sample characteristics

Sample size

The sample size of included studies ranged from 11 to 475 (Figure 9). Of the studies that collected quantitative data, 10 had fewer than 30 participants; 10 had 30 to 100 participants; four studies had 101 to 150; and one had over 400 participants. One study[12] did not specify its sample size. The reason for the large participant count (n=475) in one study[19] was that all students taking the English debating course throughout the university were included.

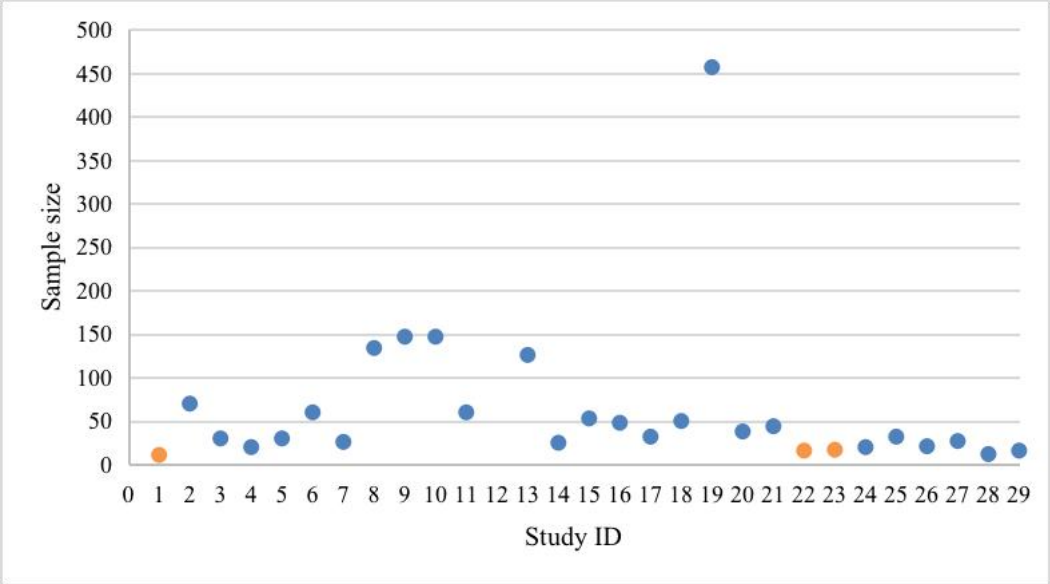

Figure 9. Sample size

Note: orange dots indicate studies that only collect qualitative data, blue dots stand for quantitative or both.

Participants' characteristics

The detailed descriptions of the participants' backgrounds, including their English proficiency, academic background, and how they were selected for intervention, if reported, can be found in Appendix H. Their main characteristics are discussed as follows.

Regarding ESL/EFL proficiency, participants generally had previous exposure to English language training before the debate intervention. Seven studies[7,8,10,13,14,18,23] clearly reported this information based on the Common European Framework of Reference for Languages (CEFR). This is a globally recognised benchmark for evaluating language proficiency which classifies language learners into six levels, namely A1 (starter/beginner), A2 (elementary/pre-intermediate), B1 (intermediate), B2 (upper intermediate), C1 (advanced), C2 (fluent/proficient) (Council of Europe, 2001). Of the eligible studies, the majority of participants ranged from B1 to B2 levels, indicating a range from intermediate to upper-intermediate proficiency. The

other nine studies[2,5,12,17,21,24,26,28,29] reported that participants were English majors, enrolled on language-related programmes, had taken English entrance examinations, and had completed ESL courses. One study[18] categorised the participants in the A1 and A2 bands, but they had been learning English for seven to 10 years. Therefore, it can be inferred that participants in eligible studies should have at least an elementary level of English as a prerequisite for participating in the debate intervention.

In terms of how the participants were selected for the intervention, most studies adopted non-probability samples (i.e. individuals are chosen based on specific criteria rather than at random; thus, not everyone has the opportunity to be included), whilst three studies[6,11,25] reported using probability samples through randomisation (i.e. a random selection of participants from a population and random allocation to the experimental or control group; thus, every member of the population has an equal selection chance). Of studies that employed non-probability samples, there were convenience samples[2,3,7,8,10,17,27] (i.e. participants were selected based on their availability or accessibility, such as using an intact class), purposive samples[1,4,16,24] (i.e. participants were selected based on specific criteria such as language proficiency), and voluntary samples[15,18,23] (i.e. participants volunteered to attend the research). See Cohen et al. (2018) for a detailed classification and discussion of sampling strategies.

4.2.5 Intervention details

This sub-section summarises the intervention process of the debate group and control group (if any). Understanding how interventions were conducted can provide invaluable pedagogical implications. Appendix I provides a detailed description of the intervention process. The data collected was highly heterogeneous, making it difficult to compare studies directly.

Despite the heterogeneity, I broadly categorised the debate process into three phases, namely, pre-debate activity, actual debate, and post-debate activity. This division is adapted from Carinanos-Ayala et al. (2021) in their systematic review of debate pedagogy amongst undergraduate nursing students. Debate formats and frequently discussed topics are also illustrated, as they are relevant to the debate preparation and the actual debate process. Hence, the following sub-sections summarise findings according to these aspects: debate format, pre-debate activity, actual debate, post-debate activity, control group (if any), and frequently discussed topics.

Debate format

The eligible studies employed various debate formats. Eight studies explicitly reported adopting established formats such as British Parliamentary debate[12,15,28,29], Asians Parliamentary debate[1], World School Debate Championship format[16], cross-examination format[26], and four-corner debate format[4]. Some used the format designed by the researcher, such as simplified debate format[19], one-on-one and two-on-two, or a combination of different formats[8,10]. Two studies conducted debate classes online with participants coming from two countries[5] and two universities[14], and one study[13] reported using computer- and mobile-assisted language learning tools to facilitate collaboration and communication in debate classes.

The common elements amongst these formats are that they all included two sides presenting arguments (one side argues for whilst the other side argues against a given topic), rebuttals and closing speeches. Their differences mainly lie in team sizes, speakers' roles, speech order and speaking duration (as discussed in Section 2.2.2).

Pre-debate activity

In the pre-debate stage, the frequently reported activities were that teachers introduced the debate format and technique, and students consulted relevant articles. Other sporadically mentioned activities included writing assignments, brainstorming,

warm-up discussion, and vocabulary learning. Students were often given one week of preparation time before the next debate session started.

Regarding team and role allocation, there were divergences. The frequently adopted team allocation method was that one team being the affirmative/government side to support the motion, while the other being the negative/opposition to refute the motion[5,8,14,19,24]. The number of members in each team was mostly two to three [8,18,19,24,26,28,29], although studies also reported four to six[5,20,23], or even 10[14], depending on the class size. Their roles in the debates were assigned based on students' own choice[4,10,17,19], random allocation[18,26,29], or by the teachers' designation[16,19].

Actual debate

13 studies reported the implementation of actual debate. Three common elements in this phase included constructive speech, rebuttals, and the summary of clashes. The debate duration for each round ranged from 25 to over 50 minutes, whilst speaking times for each participant lasted between one and over seven minutes. 10 studies[2,3,6,7,11,13,15,21,22,27] did not clearly report the information about the actual debate phase.

Figure 10 outlines the actual debate process employed by many included studies. The process can be summarised as: the first speaker on the affirmative side presents his/her arguments first, followed by the first speaker on the opposite side who gives rebuttals and present his/her arguments. Then, the second speaker on the affirmative side presents his/her rebuttals and arguments before the second speaker on the other side presents his/her rebuttals and arguments. This process is repeated based on the number of team members, until the last speaker on each side summarises the entire debate.

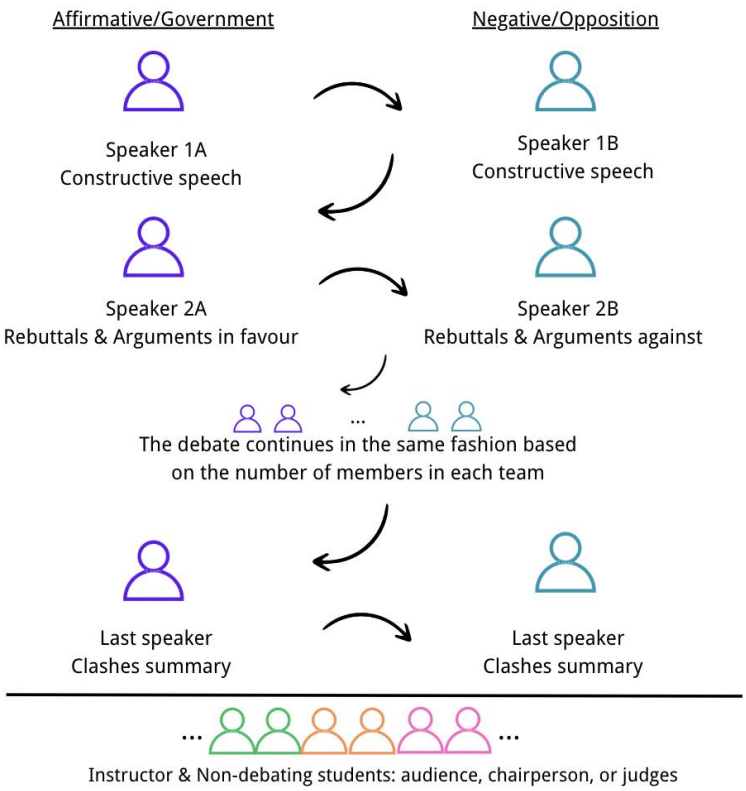

Figure 10. Actual debate process

Additionally, point of information (POI), namely individuals on the opposing team or members of the audience who have the opportunity to challenge or seek clarification on the arguments presented by the speaker on the floor, was also a frequently cited activity[1,12,16,28,29]. Non-debating students often played the role as audience members, moderators, or judges to evaluate debaters' performance[1,14,16,17,19,23,26], so that all students were involved in the classroom debates.

Post-debate activity

The third phase is the post-debate. The frequently mentioned activities were feedback issued by teachers and non-debating students (e.g., strengths and weaknesses of arguments, delivery and structure of speeches), discussion between students and teachers, self-reflection, and writing/reading assignments (e.g., journals, blogs, and argumentative essays). Six studies[2,6,7,11,21,22] did not report which debate format was

adopted, nor did they provide information regarding how the chosen debate pedagogy was implemented.

The entire intervention duration varied. Shorter interventions ranged from nine to 18 weeks, whereas longer interventions lasted from one semester to a full academic year. Three studies[4,11,28] did not report the length of their interventions.

Control group

Out of 10 studies which reported having a control group, eight[2,3,6,9,10,15,25,26] mentioned having a control group that received non-debate instruction, whilst two[11,16] reported having no treatment. Amongst them, only four studies[9,10,25,25] provided a clear description of the control group's classroom formats. These included face-to-face discussion with more reading and writing assignments compared to the experimental group[9,10], inquiry discussion[25], and a combination of activities such as discussion, presentations, film watching, charades, and role plays[26]. Others[2,3,6,15] merely mentioned employed 'traditional', 'regular' or 'conventional' teaching methods with no further elaboration Without clearly reporting this information, it is difficult to understand the relative effects of different pedagogical approaches.

Frequently discussed topics

This sub-section summarises the topics frequently discussed in the debates or speaking tasks, if reported. As the topics were presented in qualitative data form, they were analysed using Nvivo, a computer software program designed for analysing and visualising qualitative data (Phillips & Lu, 2018), for word frequency. Figure 11 depicts a word cloud representing the word frequency associated with the topics. The frequently occurring topic words primarily related to societal issues, personal well-being, lifestyle, relationships, and technology. See Appendix J for the detailed topics list and Appendix K for a complete list of topic word frequency.

Figure 11. Topic words frequency cloud

4.2.6 Methods of data collection

Outcome data were collected via various instruments. Figure 12 shows that the most widely used instrument was speaking tests, followed by questionnaires, interviews, observations, others (e.g., students' journals, reflective papers, written scripts), and field notes.

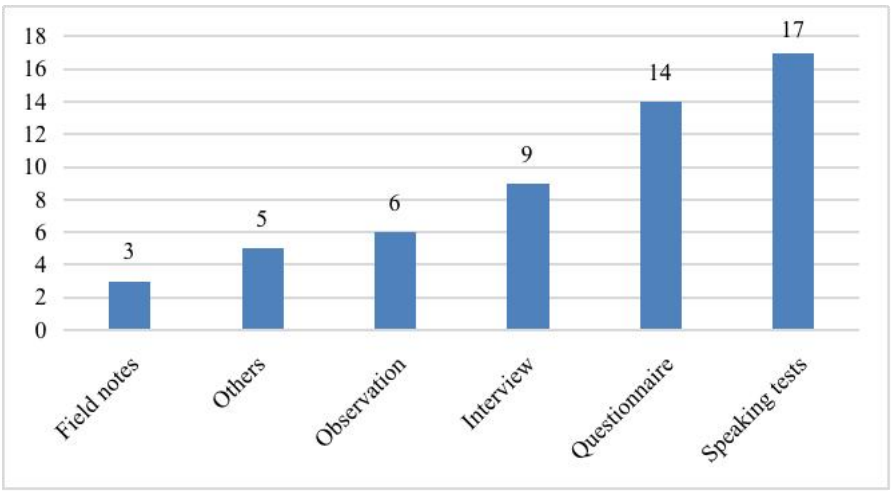

Figure 12. Methods of data collection

The speaking tests adopted were highly heterogeneous, including tests created by the researchers (e.g., opinion tasks[2,8,9,10,17,24], impromptu speeches[15,26], question-and-answer speaking tasks[2,4], and picture description[2,18]), or adapted from national or international standardised tests (e.g., oral proficiency interview designed by the American Council on the Teaching of Foreign Languages (ACTEFL)[5], and monologue using rubrics from the Student Oral Language Observation Matrix (SOLOM)[16], or non-specified speaking tests[2,6,11,12,27].

4.2.7 Reported outcomes

The outcome items reported can be broadly categorised into language and affective aspects, as discussed in Section 2.3.

Language aspects

Regarding language aspects, speaking skills reported encompassed various sub-components. They were regrouped into 14 items based on relevance to facilitate analysis and interpretation. For example, the outcomes of 'structural analysis' and 'argument complexity' in study[10] were both classified into the argumentation category due to their relevance to constructing arguments, and the frequency of argumentation appearing as a reported outcome is considered twice. Likewise, the outcomes of 'lexical richness' and 'lexical diversity' in study[15] were categorised into vocabulary, as they are both relevant to vocabulary usage. See Appendix L for the complete list of categories.

Figure 13 presents the frequency of reported outcomes in language aspects of L2 speaking. The most frequently discussed sub-skill was argumentation, namely, the ability to construct and refute arguments. This includes such elements as reasoning, rebuttal, evidence support, logic, adequacy and strength of arguments, relevance and adequacy of content, structural analysis, speculation, and hypotheses. Following this, vocabulary and fluency were also frequently reported outcomes. Other sporadically mentioned outcomes include general speaking proficiency, grammar, pronunciation,

accuracy, speech length (e.g., speaking time, number of words produced), delivery (e.g., voice, gestures/body language, enthusiasm), interaction and involvement, comprehension, complexity (e.g., lexical and syntactic complexity), coherence and cohesion, comprehensibility and clarity. Some studies[3,6,27] used broad terms such as 'speaking skills' or 'speaking capacity' without explicitly reporting which specific aspects of speaking were enhanced through their debate pedagogy.

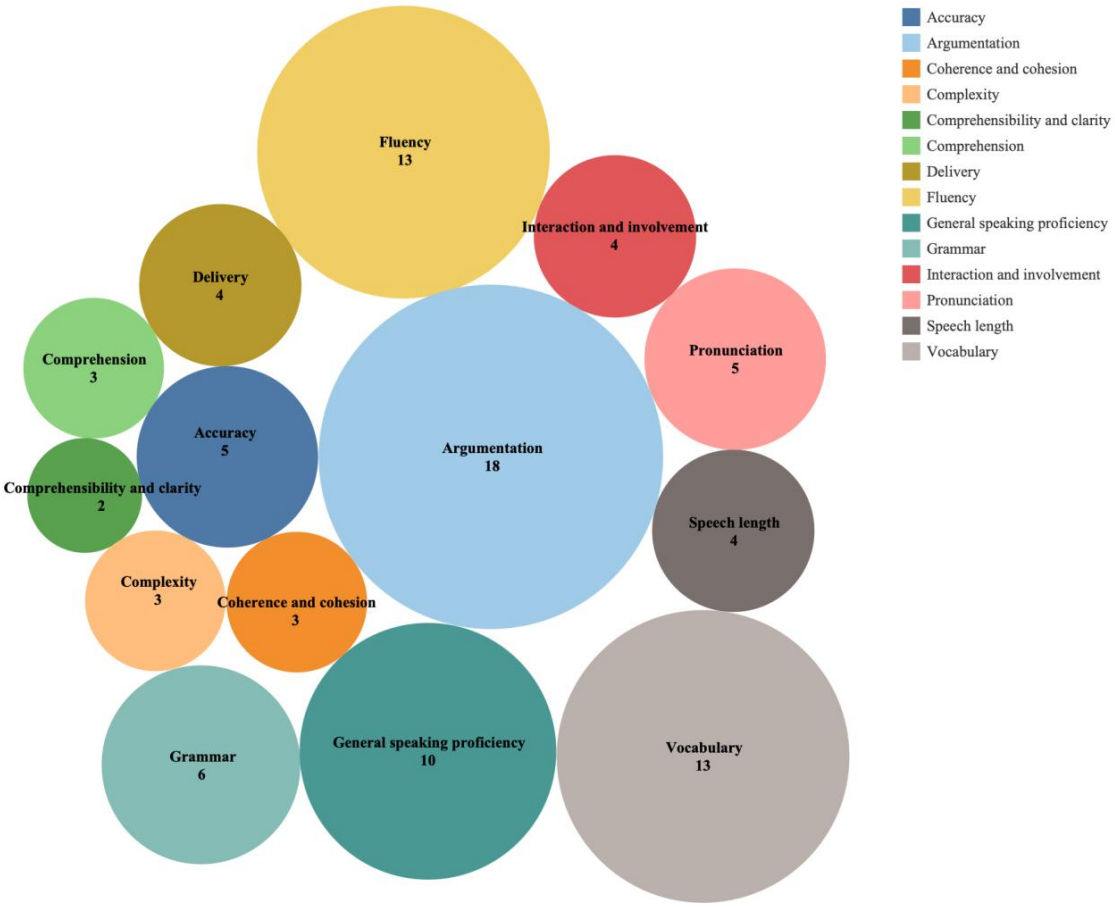

Figure 13. Frequency of reported outcome in language aspects

Affective aspects

16 studies reported affective aspects of L2 speaking, and the outcomes were relatively homogeneous (see Figure 14). These included confidence, motivation/willingness to speak English, anxiety/apprehension, and enjoyment.

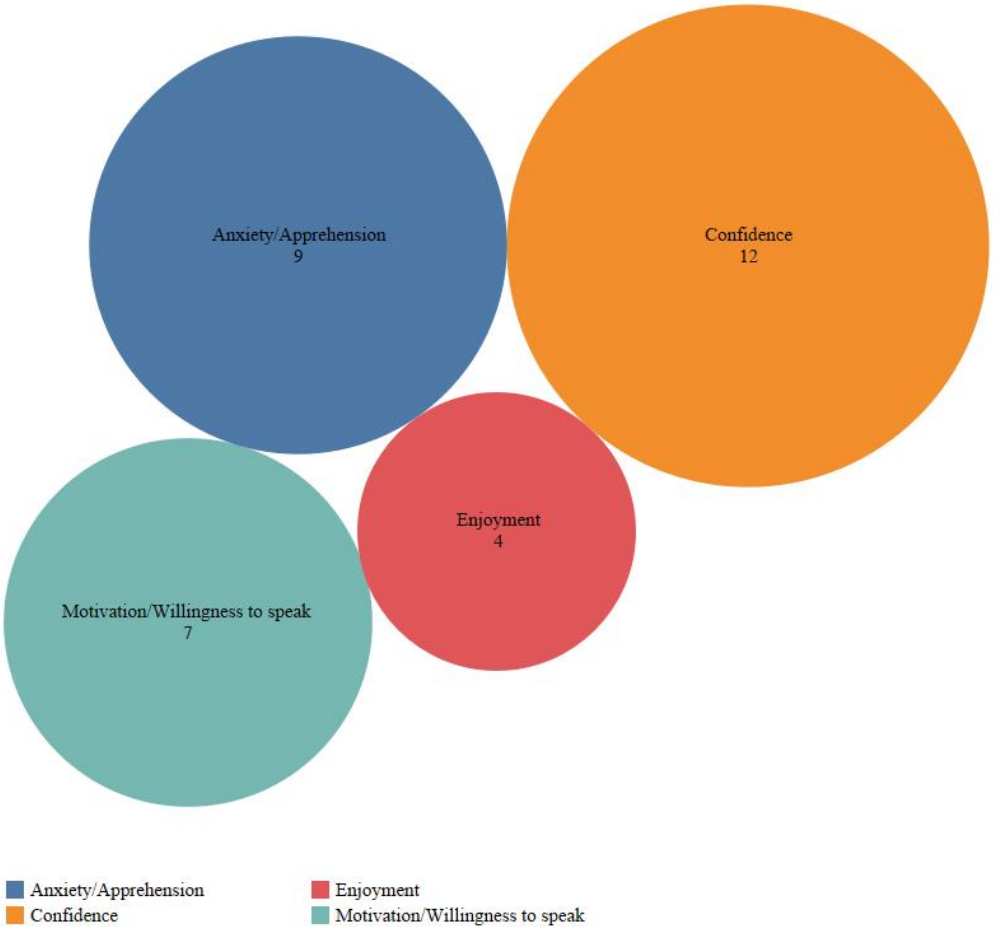

Figure 14. Frequency of reported outcome in affective aspects

Collectively, this section describes the general characteristic of included studies regarding their publication details, geographic and educational contexts, research methodology, sample descriptions, intervention specifics, measurement instruments of speaking proficiency and reported outcomes. This provides necessary contextual information for evaluating the trustworthiness of included studies in the following section.

4.3 Trustworthiness appraisal of individual studies

This section provides the trustworthiness ratings for each study, adopting the criteria and the rating guide outlined in the MMAT (see Appendix D and Appendix E). Amongst 29 studies, three were categorised as qualitative study design, three were

quantitative randomised controlled trials, 12 were quantitative non-randomised, and 11 were mixed methods.

Table 7 below presents the results of each study. As no included study employed a quantitative descriptive research design, only ratings on the aforementioned four types of study design are shown. The table features three colours, namely, green, red and yellow, which correspond to 'Yes', 'No', and 'Can't tell', respectively. A 'green' response receives a rating of one star *, while any other response receives a rating of 0. Table 7 reveals that five studies[1,8,9,10,23] received 5*, five studies[5,16,17,18,26] were 4*, four studies[13,14,15,24] were 3*, five studies[2,6,21,22,29] were 2*, six studies[3,4,11,19,25,28] were 1* and four studies[7,12,20,27] were 0. The following sub-sections further elaborate on the ratings and justifications, including screening questions, qualitative, quantitative and mixed methods studies, respectively.

Table 7. Results of the trustworthiness appraisal

Study ID	Screening		Qualitative					Quantitative randomized controlled trials					Quantitative non-randomized					Mixed methods					Overall Quality
	Clear research questions	Data suitability	Adequate design rationale	Adequate data collection	Derived findings	Substantiated interpretation	Coherence among steps	Appropriate randomization	Baseline comparability	Complete outcome data	Assessors blinding	Participants adherence	Selection bias	Appropriate intervention and outcome measurements	Complete outcome data	Accounted confounders	Intervention administration	Adequate design rationale	Integration of different components	Interpreted integration	Addressed inconsistencies	Quality of each component	
1																							5*****
2																							2**
3																							1*
4																							1*
5																							4****
6																							2**
7																							0
8																							5*****
9																							5*****
10																							5*****
11																							1*
12																							0
13																							3***
14																							3***
15																							3***
16																							4****
17																							4****
18																							4****
19																							1*
20																							0
21																							2**
22																							2**
23																							5*****
24																							3***
25																							1*
26																							4****
27																							0
28																							1*
29																							2**
	20	16	9	9	12	10	8	0	0	3	0	1	13	10	13	9	14	6	7	6	5	1	
	0	11	5	5	1	3	6	3	2	0	3	2	10	7	9	14	9	3	4	4	6	7	
	9	2	0	0	1	1	0	0	1	0	0	0	0	6	1	0	0	2	0	1	0	3	
Total	29		14					3					23					11					

4.3.1 Screening questions

Amongst 29 studies, 20 had clear research questions, whilst nine did not. For studies that failed to explicitly report this information, judging whether the collected data answered the research questions would prove difficult. Additionally, although four studies[2,11,25,28] had clear questions, the data collected may not be sufficient to answer the research questions. This includes failing to provide clear descriptions of sampling representatives, the measurement instrument used, and intervention administration, comparability between the intervention and control group, and incomplete outcome data.

4.3.2 Studies reporting qualitative data

14 of the 29 studies reported qualitative data. Eight studies[1,8,13,14,18,23,24,26] were rated as meeting all five criteria for methodological quality. Six studies that did not meet all criteria were rated as 'can't tell' or 'no' for the areas of design rationale, data collection, findings derived from the data, substantiated interpretation by data, and coherence between data sources, collection, analysis and interpretation.

Two studies[12,20] were rated 0, as they lacked clear research questions, hence the first two criteria could not be evaluated, and they did not provide detailed instrument descriptions, making it challenging to assess the following three criteria.

4.3.3 Studies reporting quantitative data

The quantitative data includes studies that adopted quantitative RCTs (n=3) and quantitative non-randomised study design (n=23). They are discussed as follows.

Randomised controlled trials

None of the three studies[6,11,25] met all criteria for methodological quality. Specifically, all three studies stated that the participants were randomly allocated into the experimental and control group without clearly describing how the randomisation

schedule was generated. Neither did they explicitly report how they ensured the intervention and control group were comparable at baseline and whether outcome assessors were blinded to the intervention provided. All three studies reported having complete outcome data, as no withdrawal/dropout rates had been reported. Participants in one study[6] were reported to have adhered to the assigned intervention, whilst two studies[11,25] failed to report this information clearly.

Quantitative non-randomised study design

Of 23 studies that employed the quantitative non-randomised study design, three studies[8,9,10] were adjudged to have met all methodological quality criteria. For the 20 studies that failed to meet one or more criteria, reasons for these ratings are explained as follows.

Regarding selection bias, 10 studies lacked sufficient information on sample representativeness. Common issues were inadequate descriptions of the target population or unrepresentative sample selection.

Issues with intervention and outcome measures were identified in 13 studies. The problems primarily stemmed from inadequate information about the content, the validity and reliability of the measurement instruments used. These make it challenging to assess whether the instruments were appropriate for the outcomes and interventions.

Concerning the completeness of outcome data, it is recommended to maintain an attrition rate of no higher than 20 percent (Petticrew & Roberts, 2006). Nine studies did not clearly report whether nearly all participants contributed to almost all measures or if only a limited proportion of participants contributed to the measures. For example, one study[18] reported a post-test for only 12 out of 50 participants due to time constraints, which resulted in a high attrition rate of 76 percent.

Regarding confounders, 14 studies did not specify if the potential confounders such as age, gender, pre-intervention test score, language exposure and proficiency were

accounted for in the research design and analysis. This may skew the results, making it difficult to determine whether the observed effects were solely due to the intervention or influenced by other variables.

In terms of the administration of the intervention, nine studies lacked transparent reporting regarding whether participants received the intended intervention as planned, whether the control group may have been indirectly exposed to the intervention between groups, and whether any unplanned co-interventions occurred within one of the groups.

4.3.4 Studies with a mixed methods design

Of 11 mixed methods studies, only one[8] met all methodological quality criteria. Two studies[18,26] received a rating of 4*, with study[18] lacked complete outcome data (only 12 of 50 finishing the speaking tests), and study[26] had potential selection bias (all participants were advanced female English learners, mostly from the English Department and one from the Engineering Department).

Eight studies received ratings of 3* or lower because they were found to have deficiencies in various areas. These include unclear rationale for the chosen research design, ineffective integration of different study components to address the research questions, inadequate interpretation of results, not addressing divergences and inconsistencies between quantitative and qualitative results, and insufficient adherence to the methodological quality criteria of different components of the study.

4.3.5 Summary of methodological quality

Overall, included studies exhibited various methodological issues. Across 29 studies, five studies received a 5* rating. This means that only 17 percent were considered to have been well-conducted as they satisfied all methodological quality criteria. Therefore, the overall methodological quality of included studies was poor.

The results of methodological quality assessment could influence the credibility of the overall results of this systematic review. Hence, they are considered when presenting the findings of included studies in the subsequent sections.

4.4 The findings of included studies

This section provides a summary of the results of individual studies. Given the limitation of space and the large number of included studies, the narrative summary for each study is provided in Appendix M. They are grouped according to their study designs, and Table 8 summarises the general reported outcomes and effects, whilst Table 9 presents studies' effect sizes if provided. These are illustrated in the following sub-sections.

4.4.1 Qualitative studies

Three qualitative studies[1,22,23] explored instructors' and students' perceptions of, and attitudes towards, using debate pedagogy to enhance L2 speaking skills through interviews and questionnaires. Some common findings in these studies included that participants perceived improvements in students' language and affective aspects of L2 speaking (detailed outcomes can be found in Appendix M). Methodologically, two[1,23] were rated 5*, whilst one[22] was rated 2* as it did not clearly report research questions, design rationale, data collection methods and the coherence among data collection, analysis and interpretation.

4.4.2 Quantitative studies

Quantitative studies in this review include non-randomised studies and RCTs. The non-randomised studies primarily employed the one-group pre-test-post-test pre-experimental design or the pre-test-post-test control group quasi-experimental design, whereas the RCTs adopted the pre-test-post-test control and experimental group design. The overall findings from quantitative studies are positive.

With regards to non-randomised studies, five[4,7,17,21,28] studies with the one-group design reported that participants improved various linguistic and affective aspects of L2 speaking after participating in classroom debates over a specified period. Similarly, seven studies[2,3,5,9,10,15,16] with the control group design revealed that the intervention group, which participated in classroom debates for a certain period, outperformed the non-debating group (e.g., four language skills training[9,10], group discussions and presentations[15], traditional method[2,3], or without treatment[16]) in several speaking skills facets. One study[5] involving Russian EFL (n=12) and Russian-as-a-foreign-language (RFL) American (n=18) participants discovered that both groups demonstrated improvement in various aspects of speaking after engaging in eight weekly debate sessions, but the difference between the pre-test and post-test was not significant, potentially due to the small sample size and short intervention.

Regarding the methodological quality amongst the non-randomised studies, only two[9,10] of the 12 studies received 5*. The main methodological issues of other studies included unclear reporting on measurement appropriateness and the control of confounding variables.

Three RCTs[6,11,25] that randomly allocated participants into the debating group and non-debating group (e.g., conventional method[6], no treatment[11], inquiry discussion[25]) showed that the debating group outperformed in various aspects of speaking than the non-debating group, except for one study[25], which found that debating triggered higher levels of anxiety in L2 speaking compared with the discussion format of instruction. Additionally, one study[6] did not specify what the 'conventional method' for the control group referred to. All three studies were found to be methodologically weak in that they failed to clearly report randomisation procedures, groups' comparability, or outcome assessors blinding.

4.4.3 Mixed methods studies

The overall findings from 11 mixed methods studies are also positive. Participants in these studies showed improvement in L2 speaking skills from pre-test to post-test, or

that the debating group performed better than the non-debating group in various aspects of L2 speaking. Interestingly, one study[18] found that after eight voluntary debate sessions, EFL learners achieved higher scores in the post-test than the pre-test, yet there was no significant difference in speaking performance and anxiety between participants with A1 and A2 English proficiency levels. Another study[26] discovered no significant difference in participants' speaking scores between the pre-test and the mid-test after six weeks of debate treatment. However, their scores were significantly improved in the post-test after 12 weeks of intervention. This suggests that the duration of debate intervention may influence the results of speaking development. Methodologically, only one[8] of the 11 studies received a 5* rating. The primary methodological issues in other studies were their lack of clarity surrounding reconciling inconsistencies between quantitative and qualitative results and ensuring each methodological component met its respective quality criteria.

4.4.4 Effect sizes

Amongst quantitative and mixed methods studies that reported quantitative data, only eight of 26 studies[2,5,6,9,10,15,16,18] reported or provided sufficient information to calculate effect sizes for their findings (see Table 8). The degree of these effect sizes (small, medium, large) was determined based on the standards proposed by Cohen (1988); namely, the effect size is small if *d* is approximately 0.2 or below, medium if *d* is approximately 0.5, and large if *d* is of approximately 0.8 or above.

As Table 9 shows, the effect sizes were medium to large in terms of overall speaking proficiency, specific language and affective aspects. For studies that reported overall speaking proficiency, three studies[2,16,18] reported large effect sizes, and two[5,6] found medium to large effect sizes. Amongst them, study[16] discovered large effect sizes in both debating and non-debating groups, but the former (d=6.13) had significantly larger effect sizes than the latter (d=1.43). Regarding specific language aspects, large effect sizes were found for quantity measures[9], cohesion[9], structural analysis[10], argument quality analysis[10], and lexical diversity[15]. Medium or medium to large effect

sizes were reported for fluency[9], syntactic complexity[9], accuracy[9], cohesion[9], and structural analysis[10]. Of two studies reporting affective aspects, one[2] found a large effect size on attitude, whilst another[18] found a medium effect size on L2 speaking anxiety. The other studies failed to provide relevant information. Without explicitly reporting effect sizes, it is difficult to make a judgement towards on the extent to which interventions have influence or make a meaningful difference.

Methodologically, amongst eight studies reporting effect sizes, two studies[9,10] were rated 5*, three[5,16,18] 4*, one[15] 3*, and two[2,6] 2*. The varying methodological quality suggests that while there were promising results concerning the efficacy of the debate interventions in ESL/EFL speaking development, caution should be exercised when interpreting these results.

Table 8. Reported outcomes and overall results

Study ID	**Reported speaking outcomes**	**Overall results**	**Quality**
Qualitative studies			
1	Language & Affective aspects	Positive	**5***
23	Language & Affective aspects	Positive	**5***
22	Language & Affective aspects	Positive	**2***
Quantitative non-randomized: the one-group pre-test-post-test pre-experimental design			
17	Language & Affective aspects	Positive	**4***
21	Language & Affective aspects	Positive, post-test > pre-test	**2***
4	Language aspects	Positive, post-test > pre-test	**1***

28	Language aspects	Positive	**1***
7	Language aspects	Positive, cycle 3 > cycle 2 > cycle 1	**0**
Quantitative non-randomized: the pre-test-post-test control group quasi-experimental design			
9	Language aspects	Positive, intervention group > control group	**5***
10	Language aspects	Positive, intervention group > control group	**5***
5	Language aspects	Neutral, both groups post-test > pre-test, but no significant difference	**4***
16	Language aspects	Positive, intervention group > control group	**4***
15	Language aspects	Positive	**3***
2	Language & Affective aspects	Positive, intervention group > control group	**2***
3	Language aspects	Positive, intervention group > control group	**1***
Quantitative randomised-controlled trials: the pre-test-post-test control and experimental group design			
6	Language aspects	Positive, intervention group > control group	**2***

11	Language aspects	Positive, intervention group > control group	**1***
25	Affective aspects	Positive, intervention group > control group, but debate triggered anxiety more	**1***
Mixed methods studies			
8	Language & Affective aspects	Positive	**5***
18	Language & Affective aspects	Positive, post-test > pre-test, but no significant different between A1 and A2 students regarding anxiety and performance	**4***
26	Language & Affective aspects	Positive, no significant difference after 6 weeks of intervention, but significant difference after 12 weeks of intervention, debating group > discussion group	**4***
13	Language & Affective aspects	Positive	**3***
14	Language & Affective aspects	Positive, post-test > pre-test	**3***
24	Language aspects	Positive, post-test > pre-test	**3***
29	Language & Affective aspects	Positive	**2***
19	Language & Affective aspects	Positive	**1***
12	Language & Affective aspects	Positive, cycle 2 > cycle 1	**0**

20	Language & Affective aspects	Positive	**0**
27	Language aspects	Positive, cycle 2 > cycle 1	**0**

Note. Detailed reported outcomes and results can be found in Appendix M for the narrative summary of included studies.

Table 9. Effect size

Study ID	**Outcome measures and effect size**	**Quality**
9	**Quantity measures** (quantity of speech, and number of words produced, both large effect sizes) **Fluency** (inverse articulation rate *d*=0.52, medium) **Syntactic complexity** (word frequency *d*=0.58, medium) **Accuracy** (error-free clauses *d*=0.71, medium to large; lexical errors *d*=0.65, medium to large) **Cohesion** (use of frame markers *d*=1.17, large; conclusion markers, *d*=0.70, medium to large; diversity of type, *d*=1.38, large; diversity of token, *d*=1.08, large)	**5***
10	**Structural analysis** (overall argument complexity *d*=0.78, medium to large; number of data *d*=1.14, large; sub-argument *d*=0.89; warrants *d*=0.65 medium to large; counterarguments *d*=0.52 medium; rebuttals *d*=0.69 medium to large) **Argument quality analysis** (overall assessment, organization, sufficiency, clarity, elaboration, relevance and persuasiveness of arguments, all large effect sizes)	**5***
5	**Oral proficiency** *d*=0.65 (medium to large)	**4***

16	**Overall speaking** *d*=6.13 (intervention group), *d*=1.43 (control group), all large effect size, but intervention group > control group	**4***
18	**L2 speaking anxiety** *d*=0.45 (medium) **L2 speaking performance** *d*=1.07 (large) **L2 speaking anxiety** (A1 vs A2) *d*=0.53 (medium) **L2 speaking performance** (A1 vs A2) *d*=1.34 (large)	**4***
15	**Lexical diversity** (Type (different words) *d*= 2.055 (intervention group), *d*=1.135 (control group); Token (total words) *d*=2.309 (intervention group), *d*=1.270 (control group), all large effect sizes, but intervention group > control group)	**3***
2	**Overall speaking** *d*=1.52, **attitudes** *d*=13.59 (large)	**2***
6	**Speaking ability** *d*=0.58 (medium)	**2***

Chapter 5. Discussion

The chapter delves into the reasons behind the study findings, and further discusses the strengths and limitations of the included studies, to critically answer two research questions of this systematic review: (1) the extent of empirical research investigating debate pedagogy, and (2) the effects of debate pedagogy on ESL/EFL speaking skills.

Sections 5.1 and 5.2 discuss studies based on their geographic and educational contexts and the implementation of debate pedagogy to answer the first research question. Sections 5.3 and 5.4 discuss studies based on ESL/EFL speaking outcomes and the methodological issues in research design to address the second question. Subsequently, Sections 5.5 and 5.6 provide the implications and limitations of this review.

5.1 Geographic and educational contexts

5.1.1 Location of studies

The rising trend in research indicates that debate pedagogy in L2 classrooms is gaining increasing attention and recognition from researchers and educators worldwide. Of the 29 included studies, 16 were conducted in Asia, with nine coming from Indonesia.

In Indonesia, for example, English is taught from primary school to university and holds significant weight in the national curriculum (Iman, 2017). Moreover, the emphasis on students' engagement and interactive speaking exercises to improve conventional teacher-led L2 pedagogy make the classroom debate a promising educational approach to foster ESL/ESL learners' communicative competence (Susanto, 2023; Wahyuni et al., 2020). This concentration suggests a keen interest in exploring debate pedagogy to develop EFL/ESL speaking skills in the region, perhaps driven by the English language's global influence and the challenges many EFL/ESL learners encounter in speaking English (as discussed in Section 2.1.2).

5.1.2 Educational contexts

Participants in included studies were either attending secondary school or in higher education with certain years of English learning experience or had attained a certain level of language proficiency, primarily B1 and above. This suggests that debates might be considered relatively challenging for younger or less proficient learners.

Theoretically, the difficulties faced by lower-performing students in participating in debates can be explained by cognitive load theory, which states that dividing attention between multiple information sources can cause cognitive overload (Sweller, 2011). When students are required to use an L2 to debate a given topic, they need to manage language and content simultaneously. This can be demanding, especially during rebuttals and the POI response process, when they need to anticipate the opponents' arguments and formulate effective responses within a limited time (Aclan & Aziz, 2015a). Compared to high-performing learners, those with less existing knowledge or lower proficiency are likely to experience a higher cognitive load in debates, as they may struggle more with basic language structures, real-time and rapid comprehension of opposing viewpoints, and the articulation of arguments (He, 2015). This highlights the need for educators and researchers to reduce debate participation barriers by controlling reasonable cognitive load, making it accessible, effective, and inclusive for learners of different proficiency levels, particularly underachieving students.

Lustigova (2011) recommended using simplified debates (e.g., step-by-step guidance and sufficient preparation before the debate, suitable and interesting debate topics, and shorter speaking time) to facilitate EFL students with intermediate and lower proficiency levels in language learning. However, Lustigova's (2011) study did not specify what these levels refer to, limiting the applicability of these findings to broader contexts.

Empirically, however, Khalil (2018) and Shan (2005) discovered that students with varying English proficiency levels did not demonstrate significant differences in the

changes in their speaking performance and anxiety levels following the debate intervention. Nonetheless, all participants in Khalil's (2018) study were advanced learners despite being categorised as having relatively higher and lower proficiency. This indicates that the distinction in their proficiency may not be significant, and its impact on speaking outcomes could also not be pronounced. Additionally, participants in Shan (2005), although categorised as A1 and A2 level, have studied English for seven to 10 years and voluntarily participated in debate intervention. Engaging in voluntary participation and acquiring knowledge through an additional language may boost learners' motivation to learn (Coyle et al., 2010). This suggests that volunteers' motivation might have also affected their better speaking outcomes rather than the direct impact of debate intervention.

5.2 Implementation of debate pedagogy

This section further discusses the implementation of debate pedagogy in different phases, namely pre-debate activity, actual debate and post-debate activity. Following this, a schedule based on the reiterative debate implementation steps identified across studies will be proposed, aiming to provide implications for integrating debate pedagogy in L2 teaching and learning practices.

5.2.1 Pre-debate activity

The importance of pre-debate preparation is widely recognised (e.g., Aclan & Aziz, 2015a; Brown, 2012; García-Sánchez & Gimeno-Sanz, 2022). Ensuring students have sufficient preparation and support could help promote their performance and confidence during the actual debate (Freeley & Steinberg, 2013; Zare & Othman, 2013, 2015), which is especially important for students of lower proficiency. They could be better prepared through various activities such as instruction on debate techniques, literature searching, reading, writing, and brainstorming with their peers. Additionally, debate topics and allocating roles in classroom debates also play an role.

Debate topics are often released a certain period before the actual debate to ensure students have sufficient time to prepare. The frequently discussed topics in included studies were societal issues, personal well-being, lifestyle, relationships, and technology. However, these were based on a small number of studies included (n=13); a broader scale of research about the topics debated by participants worldwide can be conducted to capture a more comprehensive perspective. Tailoring debate topics to align with students' interests and existing knowledge levels may encourage active participation and foster a positive language teaching and learning environment (Lee & Pulido, 2017; Zare & Othman, 2015).

Team and role allocation in debates varied. Some studies recommended smaller teams in each group to maximise speaking practice (Ang et al., 2019; Iman, 2017; Lustigova, 2011). The methods of role allocation range from self-selection and random drawing to instructors' designation, which may affect participants' engagement and satisfaction (Khalil, 2018). No single approach outperforms the others. Corbridge et al. (2013) and Eastridge (2014) favoured teacher-led assignments to encourage diverse groupings and prevent potential coalitions. Conversely, some studies found that students expressed dissatisfaction with imposed roles or random allocation, especially when forced to argue a position contradicting their beliefs (Fallahi & Haney, 2007; Randolph, 2007). They also demonstrated that they would have conducted more productive debates if they were permitted to select their teams. The diverse viewpoints indicate that the optimal strategy might differ depending on classroom objectives.

5.2.2 Actual debate

The main steps during the actual debate are taking turns for the constructive speech, rebuttals, and clash summary. POI is also a common element during actual debate, especially in parliamentary debate formats (Amiri et al., 2017).

Amongst these steps, rebuttals and POI were found to be relatively more challenging yet rewarding (Aclan & Aziz, 2015a; Nguyen, 2020). Students may prepare arguments and summaries for the side they support before the actual debate, but how the opposing team refutes their arguments and what kinds of questions the audience and opponents raise in the POI may be unprepared or unexpected (Trapp, 2016). Such situations prevent debaters from relying on prepared or memorised speeches, instead requiring them to address real-time problems by giving impromptu responses within a limited time.

This aligns with real-world situations where individuals engage in spontaneous interactions rather than relying on written notes at all times, and echoes Brown's (2004) and Ellis's (2003) viewpoint that language tasks within the classroom should reflect the authentic use of the target language in real-life contexts beyond the educational setting. This also indicates that instructors should award attention to instructing students how to make and respond to rebuttals and POI effectively.

5.2.3 Post-debate activity

In the post-debate phase, feedback provided by teachers and non-debating students, as well as reading and writing assignments, were cited by nearly half of the included studies. Feedback after a debate primarily included assessing participants' debate performances, evaluating the strengths and weaknesses of arguments, and the delivery and structure of speeches. Reading and writing assignments included journals, blogs, or argumentative essays that were related to the topic discussed during that class or the next debate session.

Engaging non-debaters in the evaluation of debaters' performance could be beneficial for both (Ang et al., 2019). Whilst debaters need to provide their arguments and rebut their opponents, non-debaters, including the teacher and student spectators, play the role of judges to evaluate the strengths and limitations of arguments and the

performance from both sides. This allows everyone involved in classroom debates to practice speaking as well as to monitor, evaluate and regulate their learning and teaching processes (Zare & Othman, 2013). Studies have found that feedback providers and receivers can experience advantages from the learning process from proper guidance (Boud & Molloy, 2013; Zhu & Carless, 2018). This implies that instructors should not only teach students debate and peer feedback techniques but also consider how to provide effective teacher feedback.

The impacts of various types of feedback in L2 teaching and learning have been extensively studied (e.g., Ellis et al., 2006; Ferguson, 2011; Fu & Li, 2022; Ghilay & Ghilay, 2015; Li et al., 2016; Sheen, 2010). However, this topic is not further discussed here, as it is not the focus of the present study. Nonetheless, as far as the researcher is aware, there is a relative scarcity of research exploring the role of feedback within the L2 classroom debate context. Such research may offer valuable insights for enhancing the quality of L2 debate courses and improving learners' speaking proficiency.

5.2.4 Proposed debate implementation schedule

Drawing from the reiterative debate implementation steps identified across studies and the preceding discussion, I propose a debate implementation schedule (Table 10). This schedule also draws inspiration from the structured guide proposed by Cariñanos-Ayala et al. (2021) for debate-based learning, implementation, and research for university nursing students in the L1 contexts. However, the proposed schedule in the present study offers more details on suggested activities in different phases than Cariñanos-Ayala et al.'s (2021). Additionally, it is designed based on the data collected from the EFL/ESL contexts, aiming to assist L2 instructors with incorporating debate pedagogy into their teaching practices.

Collectively, the integration and synergy of the different skills in the debate implementation process may contribute to the observed positive outcomes regarding

speaking proficiency (Brown & Bown, 2014; Rosas-Maldonado et al. 2020). Students actively read preparatory articles and brainstormed, which served as a basis for their written cases. The knowledge obtained from these activities, in turn, influenced their critical listening and oral production during the debates. Following the actual debate, students listened to feedback on their speaking performance from instructors and their peers, enabling them to identify areas for improvement. Additionally, they were assigned argumentative reading and writing tasks to consolidate their acquired knowledge and improve their argumentation skills (El Majidi et al., 2023). In this integration of skills, where reading serves as input for writing, writing acts as input for listening and speaking and speaking performance acts as ab output for feedback, thereby creating a virtuous cycle to develop speaking proficiency (El Majidi, 2022).

Table 10. Proposal of debate-based learning and implementation schedule

/	**Pre-debate activity**	**Actual debate**	**Post-debate activity**
Objective	- Acquiring information - Applying knowledge	- Retrieval of information - Applying knowledge using the target language	- Engagement - Self-reflection - Improvement
Students	- Read teacher-provided and self-sourced articles - Engage in brainstorming and group discussion - Draft speech outlines, argumentative essays, or notes for speeches, rebuttals, and summaries	**Debaters** - Constructive speech - Rebuttals - Clashes summary - Point of information **Non-debating students** - Moderators	**Debaters** - Listen to feedback - Group discussion - Self-reflection - Reading/writing assignments **Non-debating students** - Provide feedback

		- Evaluators - Timekeepers	- Self-reflection - Reading/writing assignments
Instructors	- Teach debate techniques, feedback methods, and useful expressions - Announce debate topics with varying preparation times - Determine team roles (self-selection/random assignment/teacher-assigned)	- Moderator - Evaluator	- Provide feedback on argument quality, language use, speech delivery, and written assignments - Assign post-reading/writing tasks - Announce debate topics and role for the next debate session, and assign pre-reading/writing tasks, if needed

5.3 ESL/EFL speaking outcomes

The included studies found that using debate as a pedagogical tool generally results in positive outcomes in ESL/EFL speaking development. The results showed improvement in both language and affective aspects of L2 speaking. The most frequently reported language outcomes included argumentation, vocabulary, and fluency, and the affective outcomes were primarily confidence, anxiety, and motivation. These findings are further discussed in the following sub-sections.

5.3.1 Language aspects

Argumentation

Argumentation is the most frequently reported outcome, which includes sub-elements such as reasoning, rebuttal, logic, and evidence support. One typical example is El Majidi et al. (2021b), which discovered that students who received 10 weeks of classroom debate intervention significantly outperformed their control counterparts, who received training in four language skills, in several measures of argument quality with large effect sizes. Notably, the intervention students did not receive any instruction on argument structure and quality, whereas the control group did. They asserted, that after being involved in classroom debates, it appears that students began to realise that persuading a critical opponent depends on the robustness and validity of their arguments.

The innate mechanism of debates may contribute to the development of argumentation skills (García-Sánchez & Gimeno-Sanz, 2022). During classroom debates, students challenge each other's reasoning. They question the validity and relevance of the presented arguments, seek further explanations, and critique flaws or inconsistencies in opponents' line of reasoning (Freeley & Steinberg, 2013). The process helps students critically evaluate arguments and better understand what constitutes high quality arguments, thereby developing their argumentation skills.

Vocabulary

Another frequently reported outcome is that participating in classroom debates boosts students' vocabulary. The debate environment provides the ideal setting for expanding participants' vocabulary (Aclan & Aziz, 2015b). One factor might be that reading, writing, speaking and listening activities during the debate process revolve around the argumentative genre (e.g., reading and writing argumentative tasks, constructing and refuting arguments) (Freeley & Steinberg, 2013). The alignment of genres across tasks may facilitate vocabulary transfer among them (Hirvela, 2016), and regular engagement in classroom debates further reinforces the memory of these new words. Moreover, acquiring extensive vocabulary in debates is advantageous and essential for constructing persuasive arguments. Participants in Mirra et al.'s (2016) study stated that they were motivated to expand their vocabulary, as it can enhance their debate speeches and help them thrive in this activity. Additionally, Aclan and Aziz (2015b) argued that learning vocabulary appears more effective when contextualised, and that students learn vocabulary in classroom debates due to the interactive nature of debates which necessitates contextualisation and promotes meaningful retrieval and utilisation of the words learned.

Fluency

Improving fluency in L2 speaking following classroom debate intervention (e.g., speaking faster and producing longer speeches) was also a frequently reported outcome. Factors such as real-time performance, preparation before tasks, and integrating reading, writing, speaking and listening activities in different debate phases may all contribute to the improved fluency (Kassem, 2021). Moreover, as discussed, the vocabulary expansion also increased the speaking pace, whilst allowing them to produce more spoken output. Having ready access to relevant vocabulary in mind reduces the need for search and accelerates the formulation process, resulting in longer, more fluent, and rapid speech (El Majidi et al., 2019a).

5.3.2 Affective aspects

The affective outcomes reported were primarily that, after taking part in classroom debates, participants felt more confident or motivated to speak English, whilst anxiety was reduced. When students participate in debates, their arguments are critiqued. This might cause them to feel anxious or evade the challenge (Freeley & Steinberg, 2013). Nevertheless, evasion is not an option in debates as they are required to uphold their position, which necessitates confidence.

Tridinanti (2018) identified a significant correlation between L2 confidence, speaking anxiety, and English-speaking performance among EFL learners. That is, students who possess higher levels of self-confidence and lower levels of anxiety are more inclined to engage in verbal interactions in the classroom. This contrasts with students with lower confidence and higher anxiety levels, who tend to avoid speaking opportunities owing to a fear of being challenged (Rasyid, 2022). Moreover, confidence is found to be negatively associated with L2 speaking anxiety (MacIntyre et al., 1998; Ozdemir & Papi, 2022), that is, an increase in students' confidence through debate intervention may be linked to reduced L2 speaking anxiety, which could encourage greater participation in speaking activities and, in turn, develop their speaking skills.

Additionally, students' motivation to succeed in debates and receive positive feedback by convincingly and confidently presenting their arguments to the adjudicator also plays a significant role (Rasyid, 2022). Such motivation impacts their willingness to present their arguments confidently, encouraging them to improve their speaking skills.

In summary, the benefits of employing debate pedagogy in developing ESL/EFL speaking are promising. However, the overall methodological quality of the research on this topic is low, which is discussed further in the next section. In addition, although many studies use similar terms to describe speaking outcomes, their measurement tools are heterogenous (see Section 4.2.6). Therefore, caution should be exercised when comparing and interpreting findings across studies.

5.4 Methodological issues in research designs

Included studies are of varying methodological quality. This section delves into the strengths and limitations of these studies that employed the qualitative and/or quantitative study design, aiming to provide implications for future research.

5.4.1 Studies reporting qualitative data

This body of research primarily reports participants' favourable perceptions and attitudes towards debate pedagogy in developing ESL/EFL speaking skills. Wilson (2022), one potentially eligible study that could not be retrieved for full-text screening due to publication issues, also collects qualitative data via interviews to investigate EFL learners' experiences of debate pedagogy as a form of teaching speaking at a private language school in the Northrn Italy.

Participants' perceptions and attitudes play a significant role in shaping their' performance (Dewaele et al., 2018; Jansen et al., 2016). Dornyei's (2002) research found that students with positive attitudes or perceptions towards a task outperformed those with relatively negative attitudes. Likewise, El Majidi et al. (2018) revealed a positive correlation between participants' attitude towards debates and their perceived benefits. Hence, researchers and instructors should investigate participants' perceptions of, and attitudes towards, debate pedagogy in language learning and align instructional methods with their expectations, encouraging involvement and improving efficacy in L2 teaching and learning.

However, whilst these studies reported the perceived benefits of debate pedagogy in enhancing L2 speaking skills, one should differentiate between the perceived effects and actual learning outcomes (Niu, 2009). The perceived effect may simply be learning potential, namely participants perceived the treatment they received has the potential to improve their task performance (El Majidi, 2022). The actual outcome, however, indicates the ability to recall or use linguistic knowledge correctly after intervention (Niu, 2009). This suggests that empirical effect studies (e.g., pre-test,

post-test, delayed-post-test) are required to substantiate the perceived benefits and to provide a more objective evaluation of classroom debates' impacts on L2 speaking development.

5.4.2 Studies reporting quantitative data

Studies reporting quantitative data primarily adopted the one-group pre-test-post-test pre-experimental design, the pre-test-post-test quasi-experimental design, and RCTs. They are critically examined in the following sub-sections.

The one-group pre-test-post-test pre-experimental design

Studies applied this type of design were mainly conducted in an intact class and without a control group for comparison. This design type has been criticised for its lack of validity and reliability (Cohen et al., 2018). Although some included studies[4,7,21] discovered statistically significant differences between the pre-test and post-test of speaking performance, it is not appropriate to attribute these differences solely to the effects of debate pedagogy due to the absence of any form of comparison. This means that participants' improved performance may be attributed to practice effects, maturation, or simply the passage of time (Knapp, 2016). With no comparison group, it is impossible to say. Additionally, using an intact class instead of randomisation limits the findings' generalisability to the broader population.

The pre-test-post-test quasi-experimental design

Studies employing this type of design had control group but lack random allocation of participants into experimental and control groups. Although many studies[2,3,9,10,15,16,26] reported that the experimental group performed better than the control group after debate intervention, the better results may have been due to greater and higher quality exposure to speaking practice rather than being a direct result of debate pedagogy. For example, the experimental group in Al-Tamimi's (2017) and He's (2015) study received more intensive speaking practice in the debates than the control group,

which adopted other activities such as discussion, presentations, charades and film watching.

More comparable studies were conducted by El Majidi et al. (2021a; 2021b). The experimental group (five intact classes) and control group (three intact classes) in these studies received the same number of sessions per week (ten weeks in total), shared the same instructors, and received the same core teaching materials; the differences were that whilst the experimental group debated once a week, the control group practiced more with four language skills and had more reading and writing assignments. Participants were tested through oral argumentative tasks, which required them to argue for or against a given topic. The studies discovered that the intervention group outperformed the control group in many aspects of speaking skills.

These studies acknowledged that the experimental group practised speaking more than the control groups, which may have improved their oral performance due to the practice effect (El Majidi et al., 2021a; 2021b). However, they also argued that other intervention-related factors, such as the way the debate tasks were designed, the debate procedure, and the intensive speaking environment, were known to affect the production of complex, accurate, coherent, and fluent speech (see Section 5.2). Nonetheless, given that these studies used intact classes, their findings may have limited generalisability beyond the specific classes examined. Moreover, they were conducted at a relatively shorter intervention period, and the lack of delayed post-test limit the understanding of the sustained effects of debate intervention.

Randomised controlled trials

RCTs are praised by many educational researchers for demonstrating 'controllability, causality (i.e. an outcome has been caused by a specific intervention) and generalisability' (Cohen et al., 2018, pp. 391-396). They are often considered the 'gold standard' for determining intervention effectiveness (Haynes et al., 2012, p. 4). Ideally, by randomly allocating students to experimental and control groups and

ensuring they are equivalent in all factors (by randomisation) except exposure to the intervention, researchers can attribute any differing results between the groups to the effects of the intervention (Morrison, 2009).

However, conducting RCTs can be costly, intricate, impractical and sometimes unfeasible, such as random sampling from large populations in numerous educational studies (Cohen et al., 2018; Torgerson, 2009). Additionally, Hage and Meeker (1988, p. 55) argued that one could not assume that 'a single cause produces an effect' as the contexts or settings might actually be the primary driver of the result rather than the intervention itself.

Concerning the included studies which employed RCTs, all had low methodological quality. For example, the intervention group in Elmiyati's (2019) study outperformed the control group (without debate treatment) in both the pre-test and post-test of speaking. However, the observed better results could be due to more speaking exposure or the experimental groups' better language proficiency before starting the experiment, not necessarily the debate intervention. Another study conducted by Dehham (2021) lacked clarity on the intervention and the speaking test used, making it difficult to discern the effects of debate intervention on speaking proficiency.

In comparison, Shamsudin et al. (2017) reported relatively more explicit contextual information. This study compared a debate group (n=16) to a discussion group (n=16) over eight weeks of intervention. Using a self-report questionnaire as the measurement instrument, both groups showed no significant differences before the treatment, but the debate group demonstrated significantly more improvement in L2 communication willingness after the intervention. However, the study's small sample size indicates that the statistical power might be weak, and the results could be simply by chance; its short intervention duration may not capture the long-term effects; and its use of a self-reported questionnaire may elicit potential response biases.

5.4.3 Summary of methodological issues

Collectively, diverse methodological issues were identified in included studies. These include:

(1) Many studies were conducted at a single-nation or single-institution level, limiting the findings' applicability to the broader contexts;

(2) Selection bias due to mainly intermediate and advanced participants, with less representation from beginner-level students, and the selection primarily from intact classes rather than randomisation, limit the findings' generalisability to broader populations;

(3) Potential practice effects, that is, the positive outcomes may be due to increased speaking practice rather than the direct impacts of debate intervention;

(4) The reliance on self-reported data in many studies makes it difficult to determine whether the positive benefits were perceived effects or actual improvements;

(5) The possible influence of motivation from voluntary participation in some studies makes it difficult to determine whether positive outcomes are solely due to the debate intervention or if motivation played a significant role;

(6) Insufficient long-term research and delayed post-tests to examine the lasting effects of debate intervention;

(7) A lack of well-designed randomised controlled trials to determine if groups' differences are due to the debate intervention or other variables.

These concerns make it challenging to draw any definitive conclusions about the direct impacts of debate pedagogy in enhancing ESL/EFL speaking skills, despite numerous positive outcomes being reported. Therefore, it is important to interpret the findings with caution and not assume that they are generalisable.

5.5 Implications of this review

The findings and discussions of included studies provide key implications for future research, and L2 teaching and learning.

Firstly, the limited study locations suggests that more supra-national and cross-institutional research is needed, if resources allow. Using technological tools such as video-conference technology (e.g., García-Sánchez, 2020; García-Sánchez & Gimeno-Sanz, 2022) might be a feasible solution for cross-regional research. This requires collaboration among L2 educators, policymakers and students.

Secondly, the predominance of participants at intermediate and advanced proficiency levels in secondary and higher education indicates that further empirical research is required to investigate the interaction between debate pedagogy and L2 speaking skills with participants from a broader range of proficiency levels and education phases. Moreover, educators should consider lowering the participation barriers and providing sufficient support for students, especially beginner-level students, making it accessible, inclusive and beneficial to students with varying capabilities. Such research and action hold academic and practical value, guiding educators and policymakers on when and for whom this educational approach might be practical.

Thirdly, a common actual debate process (Figure 10) and a debate pedagogy implementation schedule covering pre-debate, actual debate and post-debate activities (Table 10) were proposed based on the elements that were frequently discussed. These attempts have the potential to facilitate L2 instructors to integrate debate pedagogy into their teaching practices. However, to date, no research has been published investigating which debate format is the most widely adopted or effective in L2 classrooms. This suggests that educators have exhibited much discretion in tailoring their debate curriculum according to class size, lesson duration, and student capabilities.

Fourthly, failure to clearly report quality evaluation-related data, especially the research design rationale, the intervention procedure, the validity and reliability of measurement instruments used, groups' comparability at baseline, the control of confounding variables (e.g., age, gender, language proficiency), and effect sizes (if any), contributes to the low methodological quality in many included studies. This highlights the importance of transparent and explicit reporting of relevant research data.

Finally, the methodological issues identified indicate that more robust empirical research with participants of greater proficiency range, longer intervention duration and possibly RCTs are needed to provide more potent evidence on the effectiveness of debate pedagogy in L2 learning.

5.6 Limitations of this review

This review made every effort to reduce the impacts of potential bias. However, there are inevitably some limitations in this review.

Firstly, the search term 'debat*' might be too broad, resulting in significant irrelevant records at the initial screening stage. For example, in the title, *Debates in ESOL teaching and learning: Culture, communities, and classrooms* (Pitt, 2005), the term 'debates' does not refer to classroom debate but represents a scholarly discussion of 'ESOL' teaching and learning. Narrowing this keyword to 'in-class debate' or 'classroom debate' might yield more targeted search results. However, this approach may also increase the risk of overlooking potentially relevant studies. Hence, the researcher decided to utilise a broader term that captures a more comprehensive range of literature.

Secondly, despite employing various search strategies, it is possible that some relevant studies were not included. Moreover, one potentially relevant study (Wilson, 2022) could not be retrieved, yet it was considered not to affect the overall findings of the review based on the abstract provided.

Thirdly, whilst the two reviewers reached an acceptable level of inter-rater reliability, due to time constraints, the second reviewer only conducted 10 percent of the initial screening, used MMAT to review four included studies of four types of study designs, and did not involve data extraction. Hence, there may be a certain degree of subjectivity.

Finally, as the study only focus on the effects of debate pedagogy in developing ESL/EFL speaking skills, other language skills such as writing, reading and listening may also be worth investigating. Additionally, the study only discussed reported outcomes with relatively high frequency; other sporadic outcomes, such as grammar, pronunciation, accuracy, complexity, and speech delivery, could be further examined.

Chapter 6. Conclusion

This systematic review investigated the effectiveness of debate pedagogy in developing ESL/EFL speaking skills. Regarding the extent of empirical studies that explored this topic, the review included 29 studies published between 2004 and 2023. They were conducted at tertiary and secondary schools in various regions, primarily in Asia. The participants were mainly intermediate and advanced learners with certain years of English language learning experience. The implementation of debate pedagogy was categorised into pre-debate, actual debate and post-debate phases, combining speaking, reading, writing, and listening activities in different stages. The control groups (if any) primarily adopted group discussion methods or four language skills training, which tend to have less intensive and competitive speaking practices than the debating groups. The L2 speaking outcomes were measured via various instruments, especially speaking tests and questionnaires. The outcomes reported included language aspects (primarily argumentation, vocabulary, and fluency) and affective aspects (primarily confidence, motivation, and anxiety). These characteristics are highly heterogeneous, making it difficult to compare findings across studies.

Regarding the effects of debate pedagogy in developing ESL/EFL speaking skills, the overall findings reported are positive. These encompass participants' positive perceptions of, and attitudes towards, classroom debate as a pedagogical tool for L2 speaking development, the enhanced performance from the pre-test to the post-test, and that the debating group outperformed their non-debating counterparts in various language and affective aspects of L2 speaking after a certain period of classroom debate participation.

However, the trustworthiness appraisal revealed that the overall methodological quality of included studies was poor, with only 17 percent of studies adopting methods that would allow for a trustworthy contribution to answering the review questions. Numerous methodological concerns were identified across studies, making

it challenging to draw any firm conclusions on the direct impacts of debate pedagogy in developing ESL/EFL speaking skills. More well-designed research and action should be conducted to address these concerns, thereby providing more compelling evidence and ensuring that this educational approach can be beneficial to students with varying proficiency levels.

In summary, research in this field is still in its early stages, despite the widespread and growing interest. The current research agenda seems to be conducted in an ad hoc manner, or at least lacks consensus of what it should be. Greater academic and educational discussion is needed to enhance the overall quality of research in the field.

References

Abrar, M., Mukminin, A., Habibi, A., Asyrafi, F., Makmur, & Marzulina, L. (2018). "If our English isn't a language, what is it?" Indonesian EFL student teachers' challenges speaking English. *The Qualitative Report, 23*(1), 129-145.

Aclan, E. M., & Aziz, N. H. A. (2015a). Exploring parliamentary debate as a pedagogical tool to develop English communication skills in EFL/ESL classrooms. *International Journal of Applied Linguistics and English Literature, 4*(2), 1-16.

Aclan, E. M., & Aziz, N. H. A. (2015b). Why and how EFL students learn vocabulary in parliamentary debate class. *Advances in Language and Literary Studies, 6*(1), 102-113.

Adem, H., & Berkessa, M. (2022). A case study of EFL teachers' practice of teaching speaking skills vis-à-vis the principles of Communicative Language Teaching (CLT). *Cogent Education, 9*(1), 2087458.

Akerman, R. & Neale, I. (2011). *Debating the evidence : An international review of current situation and perceptions.* Reading: CfBT Education Trust.

Alasmari, A., & Ahmed, S. S. (2013). Using debate in EFL classes. *English Language Teaching, 6*(1), 147-152.

Al-Mahrooqi, R., & Tabakow, M. (2015). Effectiveness of debate in ESL/EFL context courses in the Arabian gulf: A comparison of two recent student-centered studies in Oman and in Dubai, UAE. In *Proceedings of the Second 21st Century Academic Forum Conference on Teaching, Learning, and Research in the "Just Google It" Age* (pp. 417-428).

Al-Tamimi, N. (2017). Improving speaking skill: Implications for using debatable topics in English speaking classes. *English Language Teaching and Literature.* Retrieved from https://www. researchgate. net/publication/320853097.

Amiri, F., Othman, M., & Jahedi, M. (2017). A case study of the development of an ESL learner's speaking skills through instructional debate. *International Journal of Humanities and Social Science, 7*(2), 120-126.

Anderson, S., & Mezuk, B. (2012). Participating in a policy debate program and academic achievement among at-risk adolescents in an urban public school district:1997-2007. *Journal of Adolescence, 35*(5), 1125-1236.

Ang, R. X., Chew, Q. H., Sum, M. Y., Sengupta, S., & Sim, K. (2019). Systematic review of the use of debates in health professions education–does it work?. *GMS Journal for Medical Education, 36*(4).

Asfaw, D., Alemu, T., & Yaya, E. (2021). EFL teachers' practices and implementations

of speaking activities in Ethiopian English textbooks. *Hindawi: Education Research International, Volume 2021, Article ID,* 8696278, 1–11. https://doi.org/10.1155/2021/8696278

ASReview LAB developers. (2023). ASReview LAB - A tool for AI-assisted systematic reviews (v1.2). Zenodo. https://doi.org/10.5281/zenodo.7821585

ASReview Team. (2023). ASReview: Active learning for systematic reviews. https://asreview.nl

Bahrani, B., Syatriana, E., Afdaliah, N., Syamsinar, S., & Supriadi, B. (2020). A method to push students' speaking ability in English foreign classroom. *The Asian ESP Journal, 16*(62), 5-12.

Baso, F. A. (2016). The implementation of debate technique to improve students' ability in speaking. *Exposure, 5*(2), 154-173.

Bellon, J. (2000). A research-based justification for debate across the curriculum. *Argumentation and Advocacy, 36*(3), 161-175.

Benati, A. (2017). The role of input and output tasks in grammar instruction: Theoretical, empirical and pedagogical considerations. *Studies in Second Language Learning and Teaching, 7*(3), 377-396.

Boland, A., Cherry, G., & Dickson, R. (Eds.). (2017). *Doing a systematic review: A student's guide*. Los Angeles, London.

Boud, D., Molloy, E., 2013. *Feedback in higher and professional education: Understanding it and doing it well.* Routledge, London.

Boutron, I., Page, M. J., Higgins, J. P., Altman, D. G., Lundh, A., Hróbjartsson, A., & Cochrane Bias Methods Group. (2019). Considering bias and conflicts of interest among the included studies. *Cochrane Handbook for Systematic Reviews of Interventions,* 177-204.

Branham, R. J, (1991). *Debate and critical analysis: The harmony of conflict.* Routledge.

Brown, H. D. (2004). *Principles of language learning and teaching.* USA: Longman.

Brown, N. A. (2009). Argumentation and debate in foreign language instruction: A case for the traditional classroom facilitating advanced-level language uptake. *The Modern Language Journal, 93*(4), 534-549.

Brown, N. A., Talalakina, E. V., Yakusheva, I. V., & Eggett, D. L. (2012). Argumentation and debate in the foreign language classroom: Russian and American university students collaborating through new technologies. *Russian Language Journal, 62*, 141-166.

Brown, R. S., & Nation, P. (1997). *Teaching speaking: Suggestions for the classroom.*

Retrieved from http://jalt-publications.org/old_tlt/files/97/jan/speaking.html.

Brown, T., & Bown, J. (2014). *Teaching advanced language skills through global debate: Theory and practice.* Georgetown University Press.

Bueno, A., Madrid, D., & McLaren, N. (2006). *TEFL in secondary education* (2nd ed.). Granada: Editorial Universidad de Granada.

Burns, A., & Joyce, H. (1997). Focus on speaking. National Centre for English language teaching and research, Macquarie University, Sydney. *New South Wales, Australia, 2109.*

Campo, L., Galindo-Domínguez, H., Bezanilla, M. J., Fernández-Nogueira, D., & Poblete, M. (2023). Methodologies for fostering critical thinking skills from university students' points of view. *Education Sciences, 13*(2), 132.

Carinanos-Ayala, S., Arrue, M., Zarandona, J., & Labaka, A. (2021). The use of structured debate as a teaching strategy among undergraduate nursing students: A systematic review. *Nurse Education Today, 98*, 104766.

Cenoz, J., & Santos, A. (2020). Implementing pedagogical translanguaging in trilingual schools. *System, 92*, 102273.

Cochrane Effective Practice and Organisation of Care (EPOC). (2017). *Data collection form.* Retrived May 20, 2023, from

https://epoc.cochrane.org/resources/epoc-resources-review- authors#conducting

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education* (8nd ed.). London, England; New York, New Yok : Routledge

Corbridge, S. J., Corbridge, T., Tiffen, J., & Carlucci, M. (2013). Implementing team-based learning in a nurse practitioner curriculum. *Nurse Educator, 38*(5), 202-205

Council of Europe (2001). *Common European framework for languages: learning, teaching, assessment, modern language division.* Strasbourg: Cambridge University Press.

Coyle, D., Hood, P., & Marsh, D. (2010). CLIL: *Content and language integrated learning.* Cambridge: Cambridge University Press.

Crowe, M., & Sheppard, L. (2011). A review of critical appraisal tools show they lack rigor: alternative tool structure is proposed. *Journal of Clinical Epidemiology, 64*(1), 79-89.

Crowell, A., & Kuhn, D. (2014). Developing dialogic argumentation skills: A 3-year

intervention study. *Journal of Cognition and Development, 15*(2), 363-381.

Darby, M. (2007). Debate: a teaching-learning strategy for developing competence in communication and critical thinking. *Journal of Dental Hygiene, 81*(4).

Davis, K. A., Zorwick, M. L. W., Roland, J., & Wade, M. M. (Eds.). (2016). *Using debate in the classroom: Encouraging critical thinking, communication, and collaboration.* Routledge.

Dearden, J. (2014). *English as a medium of instruction: A growing global phenomenon.* British Council.

Dehham, S. H. (2021). Enhancing Iraqi EFL preparatory school students' speaking skill by using debate technique. *Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12*(7), 3189-3196.

Desita, L. R., Supardi, I., & Suhartono, L. (2017). Improving students' speaking ability through debate technique. *Jurnal Pendidikan dan Pembelajaran Khatulistiwa (JPPK), 6*(5).

Dewaele, J. M., Witney, J., Saito, K., & Dewaele, L. (2018). Foreign language enjoyment and anxiety: The effect of teacher and learner variables. *Language Teaching Research, 22*(6), 679-697.

Dixon-Woods, M., Agarwal, S., Jones, D., Young, B., & Sutton, A. (2005). Synthesising qualitative and quantitative evidence: a review of possible methods. *Journal of Health Services Research and Policy, 10*(1), 45-53.

Dornyei, Z. (2002). The motivational basis of language learning tasks. In P. Robinson (ed.), *Individual differences and instructed language learning.* Amsterdam: John Benjamins, 137–158.

Dundes, L. (2001). Small group debates: Fostering critical thinking in oral presentations with maximal class involvement. *Teaching Sociology, 29*(2), 237-243. https://doi.org/10.2307/1318721

Eastridge, J. A. (2014). Use of collaborative testing to promote nursing student success. *Nurse educator, 39*(1), 4-5.

Eckstein, J., & Bartanen, M. (2015). British parliamentary debate and the twenty-first-century student. *Communication Studies, 66*(4), 458-473.

El Majidi, A. (2022). *Debate as a tool for L2 learning: Investigating the potential of in-class debates for second language learning and argumentation skills* (Doctoral dissertation). University Utrecht.

El Majidi A., de Graaff R., Janssen D. (2018). Students' perceived effect of in-class debates in second language learning. *The European Journal of Applied Linguistics*

and TEFL, 7, 35–57.

El Majidi A., de Graaff R., Janssen D. (2021a). Debate as a pedagogical tool for developing speaking skills in second language education. *Language Teaching Research*. Epub ahead of print October 10, 2021.

El Majidi A., de Graaff R., Janssen D. (2021b). The effects of in-class debates on argumentation skills in second language education. *System, 101*, 102576.

El Majidi, A., de Graaff, R., & Janssen, D. (2020). Debate as L2 pedagogy: The effects of debating on writing development in secondary education. *The Modern Language Journal, 104*(4), 804-821.

El Majidi, A., de Graaff, R, & Janssen, D. (2023). Debate pedagogy as a conducive environment for L2 argumentative essay writing. *Language Teaching Research,* 13621688231156998.

Ellis, R. (2003). *Task-based language learning and teaching.* UK: Oxford.

Ellis, R., & Shintani, N, (2014). *Exploring language pedagogy through second language acquisition research.* Routledge.

Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. *Studies in Second Language Acquisition, 28*(2), 339-368.

Elmiyati, E. (2019). Improving students speaking ability through debate in the classroom (A case study for students at second years students of SMAN3 Kota Bima in Academic Year 2017/2018). *International Journal of Multicultural and Multireligious Understanding, 6*(1), 229-239.

EPPI-Centre. (n.d.). *EPPI-Centre Database of education research*. Retrieved 7 May 2023, from http://eppi.ioe.ac.uk/cms

Faitaki, F., Liggins, S., & Li, J. (2023). *The effects of drama-based activities on the development of young EAL pupils' oral language: Protocol for a systematic review.* Retrieved March 25, 2023, from

https://idesr.org/#searchlibrarydetailspage

Fallahi, C. R., & Haney, C. R. (2007). Using debate in helping students discuss controversial topics. *Journal of College Teaching and Learning, 4*(10), 83-88. https://doi.org/10.19030/tlc.v4i 10.1540

Fauzan, U. (2016). Enhancing speaking ability of EFL students through debate and peer assessment. *EFL Journal, 1*(1), 49-57.

Ferguson, P. (2011). Student perceptions of quality feedback in teacher education. *Assessment and Evaluation in Higher Education, 36*(1), 51-62.

Field, K. (2017). Debating our way toward stronger thinking. *Gifted Child Today, 40*(3), 144-153.

Firmansyah, D., & Vegian, E. E. V. (2019). Improving the students' speaking skill through debate technique. *Professional Journal of English Education, 2*(6), 891-895.

Freeley, A. J., & Steinberg, D. L. (2013). *Argumentation and debate.* Cengage Learning.

Frey, B. (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation.* SAGE Publications.

Fu, M., & Li, S. (2022). The effects of immediate and delayed corrective feedback on L2 development. *Studies in Second Language Acquisition, 44*(1), 2-34.

García-Sánchez, S. (2020). Debates in English language education: A multimodal, collaborative ecosystem. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT), 10*(4), 33-50.

García-Sánchez, S., & Gimeno-Sanz, A. (2022). Telecollaborative debates in ESP: Learner perceptions and pedagogical implications. J*ournal of Teaching English for Specific and Academic Purposes*, 139-157.

Garrett, M., Schoener, L., & Hood, L. (1996). Debate: A teaching strategy to improve verbal communication and critical-thinking skills. *Nurse educator, 21*(4), 37-40.

Gass, S. M. & A. Mackey (2015). Input, interaction, and output in second language acquisition. In B. VanPatten & J. Williams (eds.), *Theories in second language acquisition* (2nd edn.). New York: Routledge, 180–206.

Gatbonton, E., & Segalowitz, N. (2005). Rethinking communicative language teaching: A focus on access to fluency. *Canadian Modern Language Review, 61*(3), 325-353.

Ghilay, Y., & Ghilay, R. (2015). FBL: Feedback based learning in higher education. *Higher Education Studies, 5*(5), 1–10. https://doi.org/10.5539/hes.v5n5p1

Goh, C. C. (2017). Research into practice: Scaffolding learning processes to improve speaking performance. *Language Teaching, 50*(2), 247-260.

Goh, C. C., & Burns, A. (2012). *Teaching speaking: A holistic approach.* Cambridge University Press.

Goodwin, J. (2003). Students' perspectives on debate exercises in content area classes. *Communication Education, 52*(2), 157-163.

Grimm P. (2010). Social desirability bias. In Sheth J., Malhotra N. (eds.), *Wiley international encyclopedia of marketing*.

Haddaway, N. R., Woodcock, P., Macura, B., & Collins, A. (2015). Making literature reviews more reliable through application of lessons from systematic reviews.

Conservation Biology, 29(6), 1596-1605.

Hage, J. and Meeker, B. F. (1988). *Social causality*. London: Unwin Hyman Ltd.

Hall, D. (2011). Debate: Innovative teaching to enhance critical thinking and communication skills in healthcare professionals. *The Internet Journal of Allied Health Sciences and Practice, 9*(3), 1-8.

Hannan, J., Berkman, B., & Meadows, C. (2012). *Introduction to public forum and congressional debate.* International Debate Education Association.

Haynes, L., Goldacre, B., & Torgerson, D. (2012). *Test, learn, adapt: Developing public policy with randomised controlled trials.* London: Cabinet Office.

He, J. (2015). The effects of an argumentation and debate course on students' impromptu English speaking competence: An empirical investigation. *English Teaching and Learning, 39*(3), 79-106. DOI: 10.6330/ETL.2015.39.3.04

Hirvela, A. (2016). *Connecting reading and writing in second language writing instruction* (2nd ed.). The University of Michigan Press.

Hong Q. N., Pluye P., Fàbregues S., Bartlett G., Boardman F., Cargo M., Dagenais P., Gagnon M-P., Griffiths F., Nicolau B., O' Cathain A., Rousseau M-C., & Vedel I. (2020). *Reporting the results of the MMAT (version 2018)*. Retrieved 10 May, 2023 from

http://mixedmethodsappraisaltoolpublic.pbworks.com/w/file/fetch/140056890/Reporting%20the%20results%20of%20the%20MMAT.pdf

Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., ... & Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for information, 34*(4), 285-291.

Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., & Cargo, M. (2021). *Mixed Methods Appraisal Tool (MMAT) version 2018: User guide*. McGill.

Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., ... & Vedel, I. (2019). Improving the content validity of the mixed methods appraisal tool: a modified e-Delphi study. *Journal of Clinical Epidemiology, 111*, 49-59.

Hughes, R., & Reed, B. S. (2016). *Teaching and researching speaking.* Taylor & Francis.

Iman, J. N. (2017). Debate instruction in EFL classroom: Impacts on the critical thinking and speaking skill. *International Journal of Instruction, 10*(4), 87-108.

International Database of Education Systematic Reviews. (n.d.). About IDESR. Retrieved March 25, 2023, from https://idesr.org/#aboutuspage

Jackson, M. (1973). Debate: A neglected teaching tool. *Peabody Journal of Education, 50*(2), 150-154.

Jansen, M., Liidtke, O., & Schroeders, U. (2016). Evidence for a positive relation between interest and achievement: Examining between-person and within-person variation in five domains. *Contemporary Educational Psychology, 46*, 116-127. https://doi.Org/10.1016/j.cedpsych.2016.05.004

Jodoi, K. (2023). The correlations between parliamentary debate participation, communication competence, communication apprehension, argumentativeness, and willingness to communicate in a Japanese context. *Argumentation, 37*(1), 91-118.

Jouini, S. (2019). The role of competitive debating in skills and knowledge building among ESL learners in Tunisia. *The Journal of Teaching and Education,* 67-75.

Jung, U. O. H. (1995). Retrospective review article: Speaking-the second skill. *System, 23*(1), 107–112

Kassem, H. M. (2021). Training EFL learners on debating: Effects on their oral and written performance, ideal L2 self, and communication apprehension. *MEXTESOL Journal, 45*(4).

Kennedy, R. (2007). In-class debates: Fertile ground for active learning and the cultivation of critical thinking and oral communication skills. *International Journal of Teaching and Learning in Higher Education, 19*(2), 183-190.

Kennedy, R. (2009). The power of in-class debates. *Active Learning in Higher Education, 10*(3), 225-236. https://doi.org/10.1177/1469787409343186

Khalil, E. F. A. (2018). *The effects of debate instruction on Turkish EFL learners' L2 speaking anxiety, L2 speaking performance, and L2 writing performance* (Master's dissertation). Bahçeşehir Üniversitesi, İstanbul.

Khan, S. A., Omar, H., Babar, M. G., & Toh, C. G. (2012). Utilization of debate as an educational tool to learn health economics for dental students in Malaysia. *Journal of Dental Education, 76*(12), 1675-1683.

Knowles, L. (2023). Who to Argue with: Japanese EFL Students' Preference for Student-teacher or Student-student Debate Format. *The Educational Review, USA, 7*(2), 152-160.

Lado, R. (1995). *Language teaching: A scientific approach.* New York, NY: McGraw Hill Book Company.

Lampkin, S. J., Collins, C., Danison, R., & Lewis, M. (2015). Active learning through a debate series in a first-year pharmacy self-care course. *American Journal of Pharmaceutical Education, 79*(2).

Lee, S., & Pulido, D. (2017). The impact of topic interest, L2 proficiency, and gender on EFL incidental vocabulary acquisition through reading. *Language Teaching Research, 21*(1), 118-135.

Levelt, W. J. (1995). The ability to speak: From intentions to spoken words. *European Review, 3*(1), 13-23.

Li, S., Zhu, Y., & Ellis, R. (2016). The effects of the timing of corrective feedback on the acquisition of a new linguistic structure. *The Modern Language Journal, 100*(1), 276-295.

Liao, G. (2009). Improvement of Speaking Ability through Interrelated Skills. *English Language Teaching, 2*(3), 11-14.

Lieb, M. (2007). Teaching debate skills to intermediate and lower level EFL students. *Teachers Helping Teachers Learning Applications for a Developing Nation in the 21st Century*, 72.

Litan, R. (2020). *Resolved: Debate can revolutionize education and help save our democracy*. Brookings Institution Press.

Liu, X. (2015). The challenges and opportunities for Chinese overseas postgraduates in English speaking universities. *Higher Education Studies, 5*(3), 45-57.

Loewen, S., & Sato, M. (2018). Interaction and instructed second language acquisition. *Language Teaching, 51*(3), 285-329.

Long, M. H, (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *Handbook of second language acquisition*. Academic Press, 413-468.

Long, M. H, (2018). Interaction in L2 classrooms. In J. I. Liontas (Eds.), *The TESOL encyclopedia of English language teaching*. John Wiley & Sons, Inc, 1-7. https://doi.org/10.1002/9781118784235.eelt0233

Lustigova, L. (2011). Speak your mind: Simplified debates as a learning tool at the university level. *Journal on Efficiency and Responsibility in Education and science, 4*(1), 18-30.

MacArthur, C. A., Ferretti, R. P., & Okolo, C. M. (2002). On defending controversial viewpoints: Debates of sixth graders about the desirability of early 20th-century American immigration. *Learning Disabilities Research and Practice, 17*(3), 160-172.

MacIntyre, P. D., Clément, R., Dörnyei, Z., & Noels, K. A. (1998). Conceptualizing willingness to communicate in a L2: A situational model of L2 confidence and affiliation. *The modern language journal, 82*(4), 545-562.

Mamtani, M., Scott, K. R., DeRoos, F. J., & Conlon, L. W. (2015). Assessing EM patient safety and quality improvement milestones using a novel debate format. *Western Journal of Emergency Medicine, 16*(6), 943.

Merrell, B., Calderwood, K. J., & Flores, R, (2015). The intersection of competitions and classrooms in forensics pedagogy. *Communication Studies, 66*(4), 433-447.

Merrell, B., Calderwood, K. J., & Graham, T. (2017). Debate across the disciplines: Structured classroom debates in interdisciplinary curricula. *Contemporary Argumentation & Debate, 37*, 57-74.

Mirra, N., Honoroff, B., Elgendy, S., & Pietrzak, G. (2016). Reading and writing with a public purpose: Fostering middle school students' academic and critical community literacies through debate. *Journal of Language and Literacy Education, 12*(1), 1-22.

Moher, D., Liberati A., Tetzlaff J., Altman D. G., The PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of Internal Medicine, 151*(4), 264-269.

Morrison, K. R. (2009). *Causation in educational research.* Routledge.

Nguyen, T. (2020). Introducing debate into the English classroom. *BULLETIN of Kanagawa Prefectural Institute of Language and Culture Studies, 9*, 73-93.

Niu, R. (2009). Effect of task-inherent production modes on EFL learners' focus on form. *Language Awareness, 18*(3-4), 384-402.

https://doi.org/10.1080/09658410903197256

Nobuhara, M. (2016). Analysis of debate classes and potential factors related to leading Japanese universities' education. *BESSH-2016 Full Paper Proceeding, 114*(2), 49-53. Retrieved from

http://academicfora.com/wp-content/uploads/2016/04/FULL-PAPER-PROCEEDING-TKS-436-124-1.pdf

Noor, S., & Tajik, O. (2022). Empowering speaking skill through debate: The case of Afghan EFL learners at Herat university. *International Journal of Multidisciplinary:Applied Business and Education Research, 3*(12), 2631-2640.

Nováková, Z. (2016). How speaking is taught - overview of approaches to teaching speaking. In Zuzana STRAKOVÁ (Ed). English Matters VII (a collection of papers) [online]. Prešov: Prešovská univerzita v Prešov, Filozofická fakulta. s. 79-92. Retrived June 20, 2023 from

https://www.pulib.sk/web/kniznica/elpub/dokument/Strakova3/subor/hlavicka.pd

Omelicheva, MY. (2006). Global politics on trial: Using educational debate for teaching controversies of world affairs. *International Studies Perspectives, 7*(2), 172-186.

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan - a web and mobile app for systematic reviews. *Systematic reviews, 5*(1), 1-10.

Ozdemir, E., & Papi, M. (2022). Mindsets as sources of L2 speaking anxiety and self-confidence: The case of international teaching assistants in the US. *Innovation in Language Learning and Teaching, 16*(3), 234-248.

O'Mahoney, A. (2015). The ancient art of 'debating' and its usefulness in the foreign language classroom: the case of the French Debating Association. *Journal of Language and Cultural Education, 3*(1), 144-151.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *International Journal of Surgery, 88,* 105906.

Pearson, W. S. (2019). Critical perspectives on the IELTS test. *ELT Journal, 73*(2), 197-206.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide.* Blackwell Publishing.

Phillips, M., & Lu, J. (2018). A quick look at NVivo. *Journal of Electronic Resources Librarianship, 30*(2), 104-106.

Pica, T. (1996). The essential role of negotiation in the second language classroom. *JALT Journal, 18*(2) 241-68.

Pitt, K. (2005). Debates in ESOL teaching and learning: Culture, communities and classrooms. Psychology Press.

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., ... & Duffy, S. (2006). Guidance on the conduct of narrative synthesis in systematic reviews. A product from the ESRC methods programme *Version, 1*(1), b92.

Purnama, I. (2019). *Integrating British parliamentary debating system to build student's self-confidence in EFL speaking class* (Master's dissertation). UIN Sunan Gunung Djati.

Putri, A. A., & Rodliyah, R. S. (2021). EFL Students' perception on the use of debate in speaking classroom. In *The Thirteenth Conference on Applied Linguistics (CONAPLIN 2020),* 21-29. Atlantis Press.

https://doi.org/10.2991/assehr.k.210427.004

Rahim, A., & Quraishi, P. (2019). Investigating EFL students' poor speaking skills at Kandahar University. *American International Journal of Education and Linguistics Research, 2*(2), 1-9.

Randolph, D. S. (2007). Student perceptions of the use of debate as a teaching strategy in

the allied health professions. *Journal of Allied Health, 36*(1), 13-29.

Rao, P. S. (2019a). The importance of speaking skills in English classrooms. *Alford Council of International English & Literature Journal (ACIELJ), 2*(2), 6-18.

Rao, P. S. (2019b). The role of English as a global language. *Research Journal of English, 4*(1), 65-79.

Rasyid, F. (2022). Pathway of involvement in debate to critical thinking skills, confidence, and speaking ability. In *Proceedings International Conference on English Language Teaching (ICELT 2022)*, 365-374. Tadris Tadris Bahasa Inggris IAIN Kediri.

Richards, J. (2008). *Teaching listening and speaking from theory to practice.* Cambridge University Press.

Richards, J. C. (2006). Developing classroom speaking activities: From theory to practice. *Guidelines-Singapore-Periodical for Classroom Language Teachers Then Magazine for Language Teachers, 28*(2), 3.

Richards, J. C., Richards, J. C., & Renandya, W. A. (Eds.). (2002). *Methodology in language teaching: An anthology of current practice.* Cambridge university press.

Rosas-Maldonado, M., Bascuñán, D., & Martin, A. (2020). Potential benefits of debates: Perceptions of EFL pre-service teachers. *Lenguas Modernas,* (55).

Rutten, K., & Soetaert, R. (2013). Rhetoric, citizenship, and cultural literacy. CLCWeb: *Comparative Literature and Culture, 15*(3), 1-9.

Sabbah, S. R. M. (2015). *The effectiveness of using debates in developing speaking skills among English majors at University of Palestine* (Master's dissertation). Al-Azhar University, Gaza.

Santhanasamy, C., & Yunus, M. M. (2022). A Systematic Review of Flipped Learning Approach in Improving Speaking Skills. *European Journal of Educational Research, 11*(1), 127-139.

Scannapieco, F. A, (1997). Formal debate: An active learning strategy. *Journal of Dental Education, 61*(12), 955-961.

Schneider, E. W. (2020). *English around the world: An introduction.* Cambridge University Press.

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., ... & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *Bmj,* 349.

Shamsudin, M., Othman, M., Jahedi, M., & Aralas, D. (2017). Enhancing English learners' willingness to communicate through debate and philosophy inquiry

discussion. *English Language Teaching, 10*(8), 145-152.

Shan, L. (2005). *The effect of debate on oral communication skills among university students in Taiwan: A case study* (Master's dissertation). National Tsing Hue University.

Sheen, Y. (2010). Introduction: The role of oral and written corrective feedback in SLA. *Studies in Second Language Acquisition, 32*(2), 169-179.

Stewart, S., & Stewart, W. (2014). Teaching bayesian statistics to undergraduate students through debates. *Innovations in Education and Teaching International, 51*(6), 653-663.

Stewart, T., & Pleisch, G. (1998). Developing academic language skills and fluency through debate. *The Language Teacher, 22*(10), 27-32.

Susanto, I. (2023). Speaking skill assessment through debate for university students. *English Language Education Journal (ELEJ), 2*(1), 29-43.

Swain, M. (2013). The inseparability of cognition and emotion in second language learning. *Language Teaching, 46*(2), 195-207.

Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics, 16*(3), 371-391. https://doi.Org/10.1093/applin/16.3.371

Sweller, J. (2011). Cognitive load theory. *Cognition in Education, 55*, 37.

Syahputra, M., & Chaira, S. (2020). The cultivation of students' problem-solving skill through Asian parliamentary debate system. *Getsempena English Education Journal, 7*(1), 86-101. https://doi.org/10.46244/geej.v7i1.1047

Tannen, D. (1998). *The argument culture: Moving from debate to dialogue.* Random House.

The Campbell Collaboration. (n.d.). Evidence synthesis tools for Campbell authors. Retrieved March 25, 2023, from https://www.campbellcollaboration.org/research-resources/resources.html

Thornbury, S. (2005). *How to teach speaking.* Longman.

Thornbury, S., & Slade, D. (2006). *Conversation: From description to pedagogy.* Cambridge University Press.

Tianame, M., Usman, B., & Muslem, A. (2019). The implementation of debate technique to enhance the students' speaking skill. *English Education Journal, 10*(2), 180-195.

Torgerson, C. J. (2009). Randomised controlled trials in education research: a case study

of an individually randomised pragmatic trial. *Education 3-13, 37*(4), 313-321.

Trapp, R. (2016). *Building global relations through debate*. Beijing: Foreign Language Teaching and Research Press.

Tridinanti, G. (2018). The correlation between speaking anxiety, self-confidence, and speaking achievement of undergraduate EFL students of private university in Palembang. *International Journal of Education and Literacy Studies, 6*(4), 35-39.

Tumposky, N. R. (2004). The debate debate. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 78*(2), 52-56.

University of York. (n.d.). About PROSPERO. Retrieved March 25, 2023, from https://www.crd.york.ac.uk/prospero/#aboutpage

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., & Oberski, D. (2020). *ASReview: Open source software for efficient and transparent active learning for systematic reviews.* http://arxiv.org/abs/2006.12166

van Haastrecht, M., Sarhan, I., Yigit Ozkan, B., Brinkhuis, M., & Spruit, M. (2021). SYMBALS: A systematic review methodology blending active learning and snowballing. *Frontiers in Research Metrics and Analytics, 6*, 685591.

Viknesh N., & Yunus, M. M. (2021). A systematic review of digital storytelling in improving speaking skills. *Sustainability, 13*(17), 9829.

Wahyuni, S., Qamariah, H., Syahputra, M., Yusuf, Y. Q., & Gani, S. A. (2020). Challenges and solutions to develop critical thinking with the British parliamentary debate system in EFL classrooms. *International Journal of Language Studies, 14*(3).

Wahyuningsih, S., & Afandi, M. (2020). Investigating English speaking problems: Implications for speaking curriculum development in Indonesia. *European Journal of Educational Research, 9*(3), 967-977.

Wilson, B. (2022). *Debate as a method of teaching speaking in adult EFL* (Doctoral dissertation). Teesside University.

Yulia, H., & Aprilita, N. (2017). The implementation of British parliamentary debate style training to improve second semester student's speaking ability at English education study program of Baturaja University. *International Journal of Language Teaching and Education, 1*(1), 1-7.

Zare, P., & Othman, M. (2013). Classroom debate as a systematic teaching and learning approach. *World Applied Sciences Journal, 28*(11), 1506-1513.

Zare, P., & Othman, M. (2015). Students' perceptions toward using classroom debate to

develop critical thinking and oral communication ability. *Asian Social Science, 11*(9), 158.

Zawacki-Richter, O., Kerres, M., Bedenlier, S., Bond, M., & Buntins, K. (2020). *Systematic reviews in educational research: Methodology, perspectives and application.* Springer Nature.

Zhang, X., & Zeegers, M. (2010). Redefining the role of English as a foreign language in the curriculum in the global context. *Changing English, 17*(2), 177-187.

Zhu, Q., & Carless, D. (2018). Dialogue within peer feedback processes: Clarification and negotiation of meaning. *Higher Education Research and Development, 37*(4), 883-897.

Zorwick, L. W., & Wade, J. M. (2016). Enhancing civic education through the use of assigned advocacy, argumentation, and debate across the curriculum. *Communication Education, 65*(4), 434-444.

Appendix A IDESR Protocol

Title

The effectiveness of classroom debate on EFL/ESL speaking skills: Protocol for a systematic review

Review Questions

RQ1: What is the extent of empirical research investigating debate pedagogy in the contexts of ESL/EFL speaking skills?

RQ2: What are the effects of debate pedagogy on ESL/EFL speaking skills?

Rationale

A large percentage of English as a second or foreign language (EFL/ESL) learners worldwide study English to improve their speaking skills (Rao, 2019a; Richards & Renandya, 2002). In a globalised world where oral language skills are crucial, learners need to speak English fluently to have better academic and professional prospects (El Majidi et al., 2021). However, EFL/ESL speaking practice has been marginalised in many educational settings (Baker & Westrup, 2003; Goh, 2017; Thornbury, 2005). Consequently, academics and experts have investigated various ways and methods for promoting EFL/ESL learners' speaking skills in the classroom (e.g., Brown, 2009; Leong & Ahmadi, 2017; Zare & Othman, 2015).

As an ancient Greek pedagogical tool, debate has been identified as a potentially effective pedagogy for developing speaking skills (e.g., Jackson,1973; Lustigova, 2011; O'Mahoney, 2015). It is defined as 'the process of inquiry and advocacy, a way of arriving at a reasoned judgement on a proposition' (Freeley & Steinberg, 2013, p. 6). Some studies showed that participants reported improvement in their speaking abilities due to participation in debates (Zare & Othman, 2015; Al-Mahrooqi & Tabakow, 2015). Likewise, El Majidi et al. (2018) discovered that debaters regarded in-class debate to improve their speaking skills with a mean of 4.26 on a 5-point Likert scale. More recently, El Majidi et al. (2021) presented experimental evidence that classroom debate can help improve L2 speaking skills. However, to our knowledge, no systematic review has compiled all of the primary research on the effectiveness of classroom debate on ESL/EFL speaking skills. To fully understand the potential of debate as an instructional approach to improve ESL/EFL speaking, it is imperative that all studies be considered together.

Inclusion and exclusion criteria

Bibliographic information

Include 1: Include studies with a complete reference.

Exclude 1: Exclude studies with an incomplete reference.

Rationale: Without complete reference, retrieval of works is infeasible.

Date of publication

Include 2: Include studies published on any date.

Exclude 2: Do not exclude studies based on their publication date.

Rationale: To collect the full scope of research available.

Publication status

Include 3: Include peer-reviewed articles, organisation reports, master's and doctoral thesis, and conference proceedings.

Exclude 3: Do not exclude studies based on their publication status.

Rationale: To minimise the impact of publication bias on this topic, this review includes both peer-reviewed and grey literature (e.g., conference proceedings, theses, and policy statements).

Language

Include 4: Include studies published in any language.

Exclude 4: Do not exclude studies based on their written language. However, if the main text of the eligible study is written in languages other than those the reviewer read (i.e., English and Chinese), the reviewer would try to get a translation of the text or find someone who can read this language to help with data extraction.

Rationale: The review focuses on ESL/EFL learners; it is possible that eligible studies were written up in non-English languages; thus, limiting the review to studies published only in English may lead to systematic neglect of a particular body of research.

Study design

Include 5: All types of study design.

Exclude 5: Do not exclude studies based on research design, but exclude secondary research such as systematic reviews, meta-analyses, or commentaries.

Rationale: Considering the dearth of research in this field, eliminating one kind of study design runs the danger of systematically ignoring a body of work and restricting the scope of this review.

Setting

Include 6: Include typical classroom settings (e.g., primary, secondary schools, and universities) where debate is used as a pedagogy in EFL/ESL.

Exclude 6: Exclude other settings, such as debate society, debate competition or tournament, and out-of-the-classroom/school debate.

Rationale: This review focuses on classroom settings where debate is used as a pedagogical tool; findings from studies conducted in other settings, such as debate society or debate competitions or tournaments, may have different purposes and may not generalise to typical classroom settings.

Population

Include 7: Include typically developing EFL/ESL learners from primary school to graduates with no age range.

Exclude 7: Exclude non-EFL/ESL learners and non-typically developing learners in contexts where all learners/participants are non-typical development (e.g., special education needs, developmental language disorder).

Rationale: This review focuses on the typically developing ESL/EFL population; thus, findings from non-EFL/ESL speakers and non-typically developing learners will not be included.

Intervention

Include 8: Include studies using classroom debate as an intervention.

Exclude 8: Exclude studies with no particular emphasis on classroom debate.

Rationale: This review focuses on the effects of using classroom debate as a pedagogical tool; thus, the intervention should include classroom debate.

Outcome

Include 9: Include studies that report quantitative (e.g., pre-post test scores) and/or qualitative data (e.g., attitude, perceptions, and experiences). Include outcome measures related to speaking proficiency in the target language, such as speech quantity (e.g., number of words, speaking time), fluency, accuracy, lexical complexity, syntactic complexity, pronunciation, argumentation, confidence, anxiety, and motivation.

Exclude 9: Exclude studies that only describe debate pedagogy or are thought pieces/theoretical discussions about the potential benefits of debate pedagogy – i.e., that do not report empirical outcome data on the effects of debate pedagogy - (qualitative or quantitative).

Rationale: Concrete data is essential for the synthesis of findings in this field of literature.

Information Sources

The following sources will be assessed electronically with a subscription from the University of Oxford's Bodleian Library:

Education: British Education Index, Australian Education Index

Psychology: PsycINFO

Multidisciplinary: ProQuest Social Science Premium Collection, Scopus, Web of Science Core Collection

Grey literature: ProQuest Dissertation & Theses Global, EThOS, OpenGrey

Search Strategy

An experienced librarian at the University of Oxford's Department of Education was consulted to formulate the initial search. There are three essential elements of the search string based on the review question, including:

Participants: EFL/ESL learners

Intervention: debate as a pedagogy

Outcome: speaking skills

Database search options were chosen, where available, to limit the search to searching abstracts, keywords and titles. To account for the variety of terminology used in various educational contexts, several analogous labels were included within each category. Different terms within each category were connected with the operator 'OR', and each field was joined with 'AND.'

Therefore, the Boolean search string will be, for example:

EFL OR ESL OR EAL OR ELL OR ELT OR English* OR L2 OR "second language*" OR FL OR "foreign language*"

AND

debat*

AND

speak* OR spoke* OR communicat* OR oral* OR speech* OR discuss* OR express* OR argument* OR fluency OR accuracy OR complexity OR lexical OR syntactic OR pronunciation OR confidence OR anxiety OR motivation

Additional searches will be conducted on Google to find data published by language education organisations (e.g. British Council) who file reports that may not be published in peer-reviewed journals.

Data Management

After completing all database searches, bibliographic information and abstracts will be retrieved as RIS files and uploaded to Rayyan, a programme for managing systematic reviews (Ouzzani et al., 2016) to eliminate duplicates. The remaining data will be uploaded to ASReview, an open-source machine learning-aided pipeline for systematic reviews (van Haastrecht et al., 2021), for the title and abstract screening.

After the first screening of titles, keywords and abstracts, identified full text will be downloaded and organised in Zotero. If a study meets the inclusion criteria, the investigator will extract data and record it in a Microsoft Excel file. All evaluations will be performed on a MacBook Pro running Mac OS Monterey.

Selection Process

After the completion of all searches, duplicates will be deleted in Rayyan. The remaining titles and abstracts will be reviewed in the ASReview against the eligibility criteria and tagged for full-text review. Abstracts must not expressly breach one of the exclusion criteria to be selected for full-text screening and further study. If a given study breaches a criterion, the screening process will end for that report, and the corresponding criterion will be highlighted.

After the initial screening, full texts will be retrieved. In cases where full texts are unavailable online or through the Bodleian Library, the investigator of this review will contact the study's primary author to request a copy. Only studies that explicitly satisfy all inclusion criteria will be chosen for the review.

To ensure the reliability of the investigator's screening, a second reviewer acquainted with applied linguistics will be invited to conduct a dual-reviewer blind screening. The second reviewer will initially evaluate 10% of abstracts due to time and resource constraints. During the screening process, neither reviewer will be able to examine the judgements of the other due to the blind screening capability of Rayyan. The two reviewers will meet to discuss any disparities and reach a consensus on how to address them. A Kappa score 0.7 is generally accepted in educational studies (Frey, 2018). If this threshold is not fulfilled, the reviewers will convene to discuss and resolve any disputes by consensus, and the second reviewer will then evaluate 10% more abstracts. A third reviewer will be consulted if a dispute cannot be addressed via dialogue.

Data Collection Process

Before doing the searches, a data extraction form has been created expressly for this review. The Cochrane data extraction form will be used as a model since it is adaptable and suitable for various study designs (Cochrane Effective Practice and Organisation of Care, 2017).

Once a study fulfils the inclusion criteria, the primary investigator will extract data and record it in a Microsoft Excel file. If a report has incomplete data, the primary investigator will contact the study's main author to collect the necessary information.

The second reviewer will extract data from the documents they read as well. Extracted data from the second reviewer will be compared to data from the primary reviewer, and disagreements will be resolved through discussion.

Data Items

General

Date form completed, reference citation, study author contact details, publication type, document source

Study overview

Research questions, research design, data type, location, and language of publication

Participants

Educational phase, the number of participants, population description including age, gender, language proficiency, and other relevant sociodemographic information

Intervention

Description of the intervention (e.g., details of implementation), length of intervention, control groups, sampling strategy, class grouping, baseline imbalances, attrition

Outcomes

Outcome measurement, outcome type, descriptive outcomes, effect sizes

Risk of bias/trustworthiness of individual studies

As it is anticipated that searches will bring up studies conducted using a range of research designs, the Mixed Methods Appraisal Tool (MMAT, Hong et al., 2018) will be used to evaluate each study's trustworthiness. In the MMAT, the research quality can be assessed in five methodology categories: qualitative research, randomised controlled trials, non-randomised studies, quantitative descriptive studies, and mixed methods studies (Hong et al., 2019). Across the five design categories listed above, the MMAT scores study five design-specific quality criteria with 'yes,' 'no,' or 'can't tell,' as well as a remark box to explain why a rating was provided. If a study adopts a mixed-methods strategy, multiple categories will be assessed.

The creator of MMAT advised in version 2018 not to present an overall numerical score ranging from 0 to 5 to report MMAT results as this approach does not highlight particular problems in a piece of research and raises concerns about the purpose of the quality appraisal. However, this suggestion may present difficulties for researchers when reporting MMAT results. Hong et al. (2020) offer an alternative scoring suggestion based on the previous version of MMAT. To be specific, as there are five criteria for each category of study design (see Appendix E), an overall quality score for an eligible study can be calculated using descriptors such as stars (*) or percentages (%), that is:

5***** or 100% quality criteria met

4 **** or 80% quality criteria met

3 *** or 60% quality criteria met

2 ** or 40% quality criteria met

1 * or 20% quality criteria met

For mixed methods studies, since there are 15 criteria to evaluate instead of five, the overall quality is determined by the lowest score among the study components, that is:

20% (*) when the score for the qualitative component (QUAL), quantitative component (QUAN), or mixed methods component (MM) is 1.

40% (**) when the score for QUAL, QUAN, or MM is 2.

60% (***) when the score for QUAL, QUAN, or MM is 3.

80% (****) when the scores for QUAL, QUAN, and MM are all 4.

100% (*****) when the score for QUAL, QUAN, or MM is 5.

Hong et al. (2020) argued that it is crucial to be transparent and clearly describe how the appraisal results were interpreted and used in the review. Given that critical appraisal involves making judgments, it is recommended to have at least two reviewers independently participate in the appraisal process (Hong et al., 2018). To uphold the reliability of the assessment, the second reviewer will conduct at least 10% of the risk of bias assessment.

Data Synthesis

If the included studies are consistent in the research design, populations, interventions, and outcome measures, a meta-analysis on the effect sizes of their results will be conducted using SPSS. Hedge's g (corrected effect size), as opposed to Cohen's d, will be used to compute effect sizes because it is more robust when the heterogeneity is high among studies regarding variances and sample sizes (Durlak, 2009).

However, a narrative synthesis will be undertaken if methodological discrepancy prevents a statistical synthesis or if there is insufficient data with comparable outcomes. The general framework suggested by Popay et al. (2006) for the narrative synthesis of heterogeneous literature will be followed, which includes four main elements:

1. Developing a theory of how the intervention works, why and for whom
2. Developing a preliminary synthesis of findings of included studies
3. Exploring relationships in the data
4. Assessing the robustness of the synthesis

In this approach to narrative synthesis, the textual descriptions of primary research will be organised into thematic categories based on the nature of the review question.

Meta-biases

To reduce publication bias, this review includes searches of grey literature containing theses and other unpublished articles that satisfy the inclusion criteria. To capture recent

conference evidence that has not yet been published in full-text format, searches will also include non-full-text format searches. The primary author will be contacted if it is determined that the abstracts fit the criteria for inclusion to gather as much information as possible about the study's design and outcomes.

To minimise language bias, this review will not omit studies based on the language of publication. It will cover searches in both English and Chinese (since the review team comprises fluent speakers of these languages). Any abstracts published in additional languages will be translated into Google Translate and, if feasible, discussed with a fluent speaker of that language to determine their eligibility for inclusion. It is intended that the multilingual feature of this review process would lessen language bias, which occurs when studies with only positive results are published in English journals, while studies with null or negative results are more likely to be published in non-English publications (Boland et al., 2017).

Confidence in cumulative evidence

The strength of the cumulative evidence in this research area will be assessed based on the CASP checklist appraisal tool.

https://casp-uk.net/images/checklist/documents/CASP-Systematic-Review Checklist/CASP-Systematic-Review-Checklist-2018_fillable-form.pdf

Appendix B Preliminary screening process using ASReview

Figure B1 provides an overview of the screening process using ASReview. The symbols within Figure B1 indicate whether a human or a computer carries out the activity or if both options are accessible.

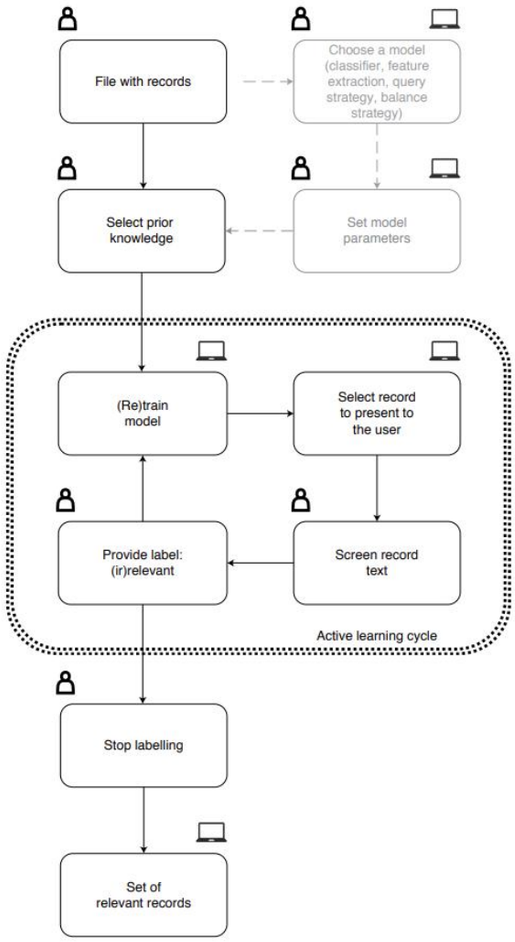

Figure B1 Machine learning-aided pipeline used in ASReview

(van de Schoot et al., 2021, p. 127)

As Figure B1 shows, the reviewer searches for, and imports metadata files containing titles and abstracts into the software. They then select prototypical reports (i.e. reports that they know meet all inclusion criteria), referred to in the schematic as 'prior knowledge' to train the first model. The software then sorts the remaining records in order of best fit based on the 'prior knowledge' examples and presents the first record for screening. The reviewer chooses whether to include or exclude this record, then the next record is presented. Once again, the reviewer decides whether to include or exclude this records, and so on. Each time a decision is made, the software adds this decision to its 'prior knowledge' and reshuffles the corpus of records, aiming to move what it regards as the best fit articles to the top of the list. This process continues iteratively throughout the screening. Researchers screen new records, labelling them as relevant or irrelevant to train a new model until meeting the stopping criterion, that is, the point at which they

determine to stop the process of screening new records (van de Schoot et al., 2021). van Haastrecht et al. (2021, p. 5) suggested using the following formula to estimate the stopping point for screening:

● Suppose you have *N* papers, with an unknown quantity of *R* papers that are relevant

● We first assess some fraction *F* of *N* papers and obtain *r* relevant studies, along with *i* unrelated papers *(i + r = F* N)*

● We can then estimate *R* as: R ≈ N * (r/(r+i))

● The stopping criterion would then be to halt screening after discovering a certain number, say 0.95 × *R*, of relevant papers

● We may also use a condition that halts the procedure upon discovering *X = (F*N)* consecutive irrelevant papers if the initial set comprising an unusually high proportion of relevant papers.

To apply the aforementioned formula in practice for this systematic review:

● There are *N* = 13, 298 papers in our data set

● I tried a fraction of *F*=0.1, so *F*N*≈1,330 papers. At this point, the number of relevant papers was *r* = 74, hence the irrelevant papers *i* =1,256.

● I then estimate *R* as: 13,296* (74/(74+1,256)) ≈ 740

● The stopping criterion would then be to halt screening after discovering 0.95×740 ≈703 studies. However, this approach may seem irrational as the number of relevant records remains relatively stable after reviewing 1,330 studies (see Figure B2). This indicates that the likelihood of discovering a new relevant record beyond this point is minimal.

● To avoid the undesirable situation that arises from having an initial set consisting of an abnormally high proportion of relevant papers, I decided to stop screening after evaluating 1,330 consecutive papers (10 percent of *N*) as irrelevant. It was discovered that within this step, a total of 76 relevant records were identified, and they were selected for further examination through a full-text review (see Figure B3).

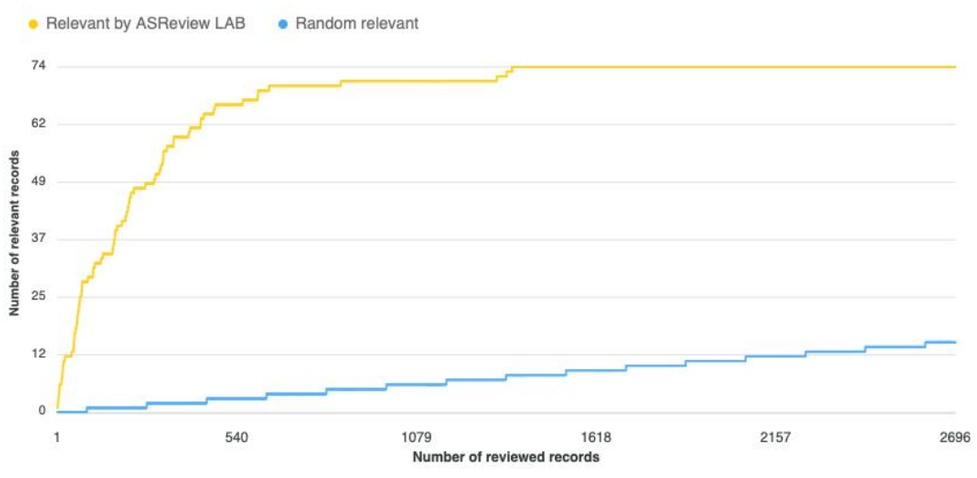

Figure B2 Trend of reviewed records

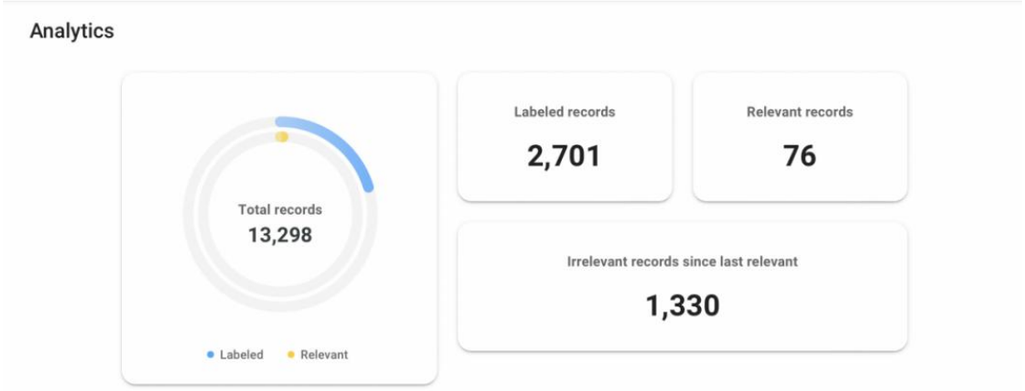

Figure B3 Relevant and irrelevant records

Appendix C Data extraction form

Category	Item	Data	Instructions
General information	Date form completed (dd/mm/yy)		Fill in the date of completion of this form.
	Reference		The full reference of the article.
	Study author contact details		Identify the contact details of the study author.
	Publication type		The publication type of the document (e.g., a thesis, peer-reviewed journal article, book chapter, etc.)
	Document source		Identify the source through which the article was specified (e.g., from an electronic database generated from the main search or through backward or forward citation searches, etc.)
	Language of publication		The language in which the article is written.
Study Overview	Research questions		Identify the research questions of the study. Put in double quotation marks ("…") if directly quoted from the study. Note as "inferred" if they were not explicitly stated in the report.
	Research design		Identify the study design for the study. For example, true experimental (e.g., RCT, quasi-experimental, pre-experimental design).
	Data type		Identify the data type of the study (e.g., quantitative, qualitative, mixed).
	Duration of study		Identity start date, end date, duration, in the format of XX month, XX weeks, XX days, where applicable. Note as "not specified" if this information is not provided in the report.
	Location of the study		What information did the authors provide about the location(s) where the study was conducted? Note as "not specified" if this information is not provided in the report.
	Education phase		Identify the school setting of the study (e.g., primary, secondary, tertiary).
	Sampling strategy		Describe how were participants selected to classroom debate sections (e.g., random sampling, convenience sampling, purposive sampling, volunteer sampling).
	Age		What is the age of the students?
	Gender		What is the gender profile of the students? Include

Participants			gender breakdowns where available.
	Other demographics information		What demographic characters do the participants have in common? If they do not have anything in common, how is this information reported? Other key information includes first language and language learning experience, for example, how is the language proficiency of participants reported in the study: with test scores, certifications, or self-reports? If no specific information is given, what proficiency level do the researchers estimate of the participants?
Intervention	Intervention procedure		Describe how the instruction is conducted in general or specifically, if any, such as how was the debating activity conducted in the L2 classroom (e.g., the types of debating format, procedure, debate topics)?
	Comparator (if any)		What method is the comparator group using? What details are given about the method? Note as "not specified" if this information is not provided in the report.
	Sample size		Fill in the total and final sample size, comprising of both number of people in the intervention group and that in the comparator. Include groups breakdowns where available.
	Length of exposure to debate pedagogy		Duration that intervention groups have been participating in debate lessons (indicate time before study and study intervention duration).
	Attrition (if any)		Did any participants leave the study? How many, and for what reasons?
Outcomes	Outcome type		What is being measured (e.g., speaking proficiency, attitude, perceptions)?
	Outcome measurement		What instruments (e.g., tests, questionnaire, interview, observation) are administered to measure learning outcomes? If the instruments were developed by the researchers, what details are given about it?
	Descriptive outcomes		Describing the outcomes reported in the study, such as characteristics (e.g., language aspects such as argumentation, vocabulary, grammar, fluency, accuracy, complexity, coherence; and/or affective aspects such as confidence, anxiety, motivation), and trends (e.g., improved, decreased, unchanged) observed in the data.
	Effect size (if any)		Record the effect size(s) as reported by the researchers, if any. Calculate using the study's information if none was given but there should have enough information.
	Conclusion (general and/or specific)		Identify the direction of the researcher's conclusions: positive, negative, neutral, or mixed.

	Critique		

Example of completed data extraction form

Category	Item	Data
General Information	Date form completed(dd/mm/yy)	19/05/2023
	Reference	El Majidi, A., de Graaff, R., & Janssen, D. (2021a). Debate as a pedagogical tool for developing speaking skills in second language education. *Language Teaching Research*, 13621688211050619.
	Study author contact details	Abid el Majidi, Utrecht University, Trans 10, 3512 JK, Utrecht, Netherlands
	Publication type	Peer-reviewed journal article
	Document source	Main search from an electronic database
	Language of publication	English
Study overview	Research questions	"What are the effects of debate pedagogy on different aspects of L2 speaking proficiency, including speech quantity (i.e. number of words, speaking time), fluency, syntactic complexity, lexical complexity, accuracy and cohesion in secondary education?"
	Study design	Pre-test-post-test control group quasi-experimental design
	Data type	Quantitative
	Duration of study	
	Location of the study	Eight intact classes in three public secondary schools in the Netherlands, located in three urban areas.
	Time of study	2021(published)
Participants	Education phase	Secondary school
	Sampling strategy	Five classes were in their higher general secondary education track (or havo 5 in Dutch)1 (n = 89), and three classes were in their fourth year of secondary pre-university education track (or vwo 4 Dutch)1 (n =58).
	Age	Between 15 and 18.
	Gender	88 female and 59 male students
	Other demographics information	The English proficiency level (including speaking) of all classes (including havo 5 and vwo 4 classes) spanned on average B1 (the intermediate level in the Common European Framework of Reference level or CEFR) as estimated by their teachers. To ensure the comparability of the groups, both the intervention and control groups consisted of both havo 5 and vwo 4 classes.
Intervention	Intervention procedure	Students in the intervention group participated in ten debates, one per week. The topics of debates (e.g. the right to bear arms) were selected in consultation with the debating students, who received one week of preparation time for each debate. Each debate consisted of three stages: pre-debate, during-debate and post-debate. In the pre-debate stage (prior to each debate), the students received a reading assignment (article) relevant to the topic under debate and were asked to find and read one additional article. We instructed the students to summarize the articles and to write a case in which they had to defend their standpoints. During actual debates, each student presented a speech and a rebuttal (in which the arguments of opponents were addressed) and participated in a

		clash. While listening to each other, the students were instructed to note down mistakes and the words they learned from each other's contributions. We used two debate formats: debating in a group of four debaters (two students in favor and two against) and a one-to-one debating format. All debates had three phases: speech, rebuttal and clash (see Snider & Schnurer, 2006). The opinion tasks consisted of different accessible topics (e.g. smoking should be banned) (see Malloy et al., 2020). These topics were randomly assigned to students. The students were explicitly instructed to provide as many arguments as possible. Before performing the task, seven minutes were allotted to planning, and there was no time limit during the performance. In the post-debate stage, the teachers provided feedback on the students' written cases and asked them to revise and resubmit them.
	Comparator (if any)	Both groups received regular instruction consisting of activities dealing with the four language skills. the intervention students participated in one debate a week. During that session, the control students received extra regular instruction in which the four language skills were further practised. While the intervention students were involved in the debate intervention (once a week), the control students received lessons based on coursebooks (during that session) in which the four language skills were further practised. These coursebooks, which were published in the Netherlands, were the same for the intervention and control groups. In each of these coursebook lessons, one language skill is mainly targeted. However, even when speaking is not the main focus of the lesson, it often figures in many lessons, for example, to activate prior knowledge. As to speaking lessons, the control group mainly practiced traditional controlled face-to-face Discussions in addition to discussing newspaper articles they read weekly.
	Sample size	Total (n = 147) Five classes served as the intervention group (n = 96) and three as the control group (n = 51).
	Length of exposure to debate pedagogy	With the exception of one intervention class that received on average two English sessions of 50 minutes per week, the other classes received three sessions of 50 minutes. Students in the intervention group participated in ten debates, one per week.
	Attrition (if any)	No attrition reported
Outcome	Outcome type	Researchers elicited speech samples from opinion tasks which they coded in terms of measures of speech quantity, fluency, complexity, accuracy and cohesion.
	Outcome measurement	**- Speech quantity:** production time (measured in seconds), which refers to the duration of the student's speech and the total number of words produced in each speech sample. **- Fluency:** speed, breakdown and repair. • Speed fluency was operationalized as inverse articulation rate (i.e. mean duration of syllables) (speech time [excluding pauses]/the number of syllables). • As to breakdown fluency, we employed two indices: mean length of pause (number of silent pauses/speech time) and the number of filled pauses (number of filled pauses/speech time). • For repair fluency, we employed two measures: number of

		repetitions (number of repetitions/speech time) and number of repairs (number of repairs and restarts/speech time). **- Syntactic complexity:** • global complexity (i.e. number of words per AS-unit); • subordination, (i.e. mean number of clauses per AS-unit); • length (i.e. mean length of clauses). **- Lexical complexity:** lexical sophistication and lexical diversity using computational tool Coh-Metrix • measure of textual lexical diversity (MTLD); • average word length (lexical sophistication); • word frequency (lexical sophistication). **- Accuracy:** • error-free clauses (EFC); • syntactic errors per 100 words; • morphological errors per 100 words; • prepositional errors per 100 words; • lexical errors per 100 words. **- Cohesion:** • frame markers: markers that mark the sequence of arguments or counterarguments (e.g. firstly, secondly); • code glosses markers: markers that introduce an example or paraphrase (e.g. for instance, in other words); • transition markers: markers that mark additive, adversative or causal relations between clauses and paragraphs (e.g. besides, because); temporal markers and the coordinating conjunction and were excluded since they are less associated with quality (Dobbs, 2014); • conclusion markers: markers that introduce a summary or conclusions (e.g. to sum up, all in all); • markers diversity token: diversity of markers in terms of token; • markers diversity type: diversity of markers in terms of token.
	Descriptive outcomes	Table 1. Means and standard errors of outcome variables across time and condition.

Table 1. Means and standard errors of outcome variables across time and condition.

Measures	Index	Intervention (n = 96)		Control (n = 51)	
		Pretest–posttest		Pretest–posttest	
Quantity	Speech quantity	25.97 (3.08)	55.43 (3.11)	32.27 (4.06)	35.6
measures	Number of words	59.22 (7.02)	128.44 (7.09)	67.88 (9.25)	75.1
Fluency	Inverse articulation rate	0.224 (.00)	0.217 (.00)	0.230 (.00)	0.23
	Number of filled pauses	0.21 (.02)	0.18 (.02)	0.23 (.03)	0.2
	Mean length of pauses	0.63 (.02)	0.60 (.01)	0.61 (.02)	0.5
	Number of repetitions	0.06 (.01)	0.06 (.01)	0.05 (.01)	0.0
	Number of repairs	0.06 (.01)	0.06 (.01)	0.06 (.01)	0.0
Syntactic	Global complexity	13.83 (.60)	13.40 (.47)	12.80 (.69)	13.9
complexity	Subordination	1.22 (.16)	1.19 (.14)	1.09 (.19)	1.2
	Length	6.31 (.17)	6.54 (.15)	6.19 (.21)	6.4
Lexical	MTLD	52.44 (2.56)	60.18 (1.93)	55.70 (2.92)	55.9
complexity	Word frequency	3.20 (.02)	3.12 (.02)	3.18 (.02)	3.1
	Word length	4.13 (.04)	4.27 (.03)	4.21 (.05)	4.3
Accuracy	Error free clauses	0.70 (.02)	0.79 (.02)	0.76 (.03)	0.7
	Lexical errors	0.92 (.17)	0.51 (.13)	0.65 (.19)	0.8
	Syntactic errors	0.94 (.14)	0.48 (.10)	0.85 (.16)	0.5
	Morphological errors	3.00 (.29)	2.42 (.22)	2.48 (.33)	2.3
	Preposition errors	0.46 (.12)	0.25 (.09)	0.57 (.13)	0.4
Cohesion	Transition markers	3.58 (.39)	4.56 (.32)	2.75 (.45)	3.8
	Frame markers	0.20 (.14)	1.24 (.11)	0.25 (.16)	0.3
	Gloss markers	0.16 (.11)	0.54 (.08)	0.22 (.12)	0.2
	Conclusion markers	0 (.04)	0.17 (.03)	0 (.04)	
	Diversity type	1.29 (.09)	2.28 (.07)	1.22 (.10)	1.3
	Diversity token	2.71 (.21)	4.61 (.15)	2.33 (.24)	2.7

Effect size (if any)

Where statistically significant results were achieved ($p < .05$), Cohen's d effect sizes (ES) are provided. The results show that the intervention students made a significant improvement in both quantity measures: quantity of speech ($F(1, 277.886) = 31.1$; $p < .001$) and number of words produced ($F(1, 279.881) = 33.9$; $p < .001$), with large effect sizes.

With regard to fluency measures, the intervention students showed a significant improvement in inverse articulation rate ($F(1, 217.650) = 4.1$; $p = .022$), with a moderate ES (.52). The other fluency measures did not reach the level of significance.

As to syntactic complexity, no significant differences were observed between the two groups at post-test. However, as to lexical complexity, the intervention group significantly outperformed the control group in terms of word frequency ($F(1, 201) = 3.6$; $p = .030$), with a moderate ES (.58). The other two lexical complexity measures, however, fell short of significance.

With regard to accuracy, the intervention students made fewer errors after the intervention relative to the control students. We found significant differences for error-free clauses ($F(1, 194.296) = 5.5$; $p = .020$), with a moderate to large ES (.71) and for lexical errors ($F(1, 194.987) = 4.8$; $p = .015$), also with moderate to large ES (.65). No significant differences were found for morphological, syntactic and preposition errors. Note that the distribution of error types did not change greatly over the intervention period.

MLM analysis also revealed that the intervention students showed significant improvement in many aspects of cohesion compared to control students. They made significant gains in terms of the use of frame markers ($F(1, 194.886) = 15.5$; $p < .001$), with a large ES (1.17), conclusion markers ($F(1, 200) = 5.3$; $p = .011$), with a moderate to large

		ES (.70), diversity of type (F(1, 200) = 20.5; p < .001), with a large ES (1.38) and diversity, of token (F(1, 200) = 12.7; p < .001), also with a large ES (1.08). **Table 2.** Multilevel analysis results. Measures / Index / Fixed effects: Denominator df / F / Sig./one tailed / *d* / Random effects: Variance within class / Variance between classes Speech quantity: Q. Speech 277.886 31.087 .000 1.33 353.51 27.96; N. Words 279.881 33.922 .000 1.38 1856.93 144.94 Fluency: Art. rate 217.650 4.140 .022 .52 .00 .00; N. filled pauses 217.374 .529 .234 .02 .00; M. length pauses 218.788 .018 .446 .02 .00; N. repetitions 217.135 .056 .407 .00 .00; N. repairs 216.132 1.372 .122 .00 .00 Syntactic complexity: Global complexity 196.668 2.099 .075 13.64 .30; Subordination 195.500 .284 .298 .70 .05; Length 195.005 .082 .388 .78 .07 Lexical complexity: MTLD 196.592 2.163 .072 282.59 2; Word frequency 201 3.606 .030 .58 .02 .00; Word length 201 .100 .377 .09 .00 Accuracy: Error free clauses 194.296 5.494 .020 .71 .02 .00; Lexical errors 194.987 4.797 .015 .65 .95 .03; Syntactic errors 195.943 .373 .271 .83 .00; Morphological errors 194.713 .754 .193 3.31 .05; Preposition errors 200 .035 .426 .63 .00 Cohesion: Transition markers 194.511 .031 .430 5.10 .20; Frame markers 194.886 15.517 .000 1.17 .69 .02; Gloss markers 200 2.709 .051 .52 .00; Conclusion markers 200 5.344 .011 .70 .06 .00; Diversity type 200 20.513 .000 1.38 .38 .00; Diversity token 200 12.706 .000 1.08 2 .00 *Note.* numerator df = 1.
	Conclusion (general and/or specific)	Multilevel analysis results indicate that after the intervention, the intervention group produced more language which was more fluent, accurate, coherent and lexically more sophisticated relative to the control group. Specifically: - The intervention students showed substantial growth in quantitative measures of oral performance. - The participants in the intervention group showed significant improvement in the speed measure of fluency. Other aspects of fluency – breakdown fluency and repair fluency – did not improve. - The debate intervention did not lead to more syntactic complexity in the produced dis course maybe due to a 'trade-off' effect. - The intervention group outperformed the control group in two measures of lexical complexity: measure of textual lexical diversity (MTLD) and word frequency, with the latter reaching significance. - The debate intervention seems to yield some benefits with respect to accuracy. Although learners in both groups showed some improvement, the intervention group made more significant improvements in two of the accuracy indices. - In the post-test, the debaters produced speech samples that demonstrated significant growth in terms of different cohesion areas.
	Critique	As acknowledged by the authors: 1. The group that received intervention practiced speaking more than the control group, which may have improved their oral performance. While acknowledging the practice effect, we argue that additional intervention-related factors contributed to the study's findings. 2. The findings in this study were based on a single task type (i.e. opinion tasks), future studies should consider employing a wider variety of different tasks to gain deeper insights into the observed effects and their transferability.

		3. This study elicited data from two time points (pre- and post-test). A delayed post test (i.e. a third time point) would provide insights into the durability of the performance patterns observed. 4. The present research has not examined the impact of debating on pronunciation, which is an important dimension of speaking performance, although previous study (El Majidi et al., 2018) reported that they felt that their pronunciation had improved after participating in debates. Its clarity, for example, is deemed to be an important feature of comprehensibility (Suzuki & Kormos, 2020).

Appendix D Mixed Methods Appraisal Tool (MMAT, version 2018)

Category of study designs	**Methodological quality criteria**	**Responses**			
		Yes	No	Can't tell	Comments
Screening questions (for all types)	S1. Are there clear research questions?				
	S2. Do the collected data allow to address the research questions?				
	Further appraisal may not be feasible or appropriate when the answer is 'No' or 'Can't tell' to one or both screening questions.				
1. Qualitative	1.1. Is the qualitative approach appropriate to answer the research question?				
	1.2. Are the qualitative data collection methods adequate to address the research question?				
	1.3. Are the findings adequately derived from the data?				
	1.4. Is the interpretation of results sufficiently substantiated by data?				
	1.5. Is there coherence between qualitative data sources, collection, analysis and interpretation?				
2. Quantitative randomized controlled trials	2.1. Is randomization appropriately performed?				
	2.2. Are the groups comparable at baseline?				
	2.3. Are there complete outcome data?				
	2.4. Are outcome assessors blinded to the intervention provided?				
	2.5 Did the participants adhere to the assigned intervention?				
3. Quantitative non-	3.1. Are the participants representative of the target population?				
	3.2. Are measurements appropriate regarding both the outcome and intervention (or exposure)?				

randomized	3.3. Are there complete outcome data?				
	3.4. Are the confounders accounted for in the design and analysis?				
	3.5. During the study period, is the intervention administered (or exposure occurred) as intended?				
4. Quantitative descriptive	4.1. Is the sampling strategy relevant to address the research question?				
	4.2. Is the sample representative of the target population?				
	4.3. Are the measurements appropriate?				
	4.4. Is the risk of nonresponse bias low?				
	4.5. Is the statistical analysis appropriate to answer the research question?				
5. Mixed methods	5.1. Is there an adequate rationale for using a mixed methods design to address the research question?				
	5.2. Are the different components of the study effectively integrated to answer the research question?				
	5.3. Are the outputs of the integration of qualitative and quantitative components adequately interpreted?				
	5.4. Are divergences and inconsistencies between quantitative and qualitative results adequately addressed?				
	5.5. Do the different components of the study adhere to the quality criteria of each tradition of the methods involved?				

Qualitative studies

1.1. Is the qualitative approach appropriate to answer the research question?

Explanations

The qualitative approach used in a study (see non-exhaustive list on the left side of this table) should be appropriate for the research question and problem. For example, the use of a grounded theory approach should address the development of a theory and ethnography should study human cultures and societies. This criterion was considered important to add in the MMAT since there is only one category of criteria for qualitative studies (compared to three for quantitative studies).

1.2. Are the qualitative data collection methods adequate to address the research question?

Explanations

This criterion is related to data collection method, including data sources (e.g., archives, documents), used to address the research question. To judge this criterion, consider whether the method of data collection (e.g., in depth interviews and/or group interviews, and/or observations) and the form of the data (e.g., tape recording, video material, diary, photo, and/or field notes) are adequate. Also, clear justifications are needed when data collection methods are modified during the study.

1.3. Are the findings adequately derived from the data?

Explanations

This criterion is related to the data analysis used. Several data analysis methods have been developed and their use depends on the research question and qualitative approach. For example, open, axial and selective coding is often associated with grounded theory, and within- and cross-case analysis is often seen in case study.

1.4. Is the interpretation of results sufficiently substantiated by data?

Explanations

The interpretation of results should be supported by the data collected. For example, the quotes provided to justify the themes should be adequate.

1.5. Is there coherence between qualitative data sources, collection, analysis and interpretation?

Explanations

There should be clear links between data sources, collection, analysis and interpretation.

Quantitative randomized controlled trials

2.1 Is randomization appropriately performed?

Explanations

In a randomized controlled trial, the allocation of a participant (or a data collection unit, e.g., a school) into the intervention or control group is based solely on chance. Researchers should describe how the randomization schedule was generated. A simple statement such as ‘we randomly allocated’ or ‘using a randomized design’ is insufficient to judge if randomization was appropriately performed. Also, assignment that is predictable such as using odd and even record numbers or dates is not appropriate. At minimum, a simple allocation (or unrestricted

allocation) should be performed by following a predetermined plan/sequence. It is usually achieved by referring to a published list of random numbers, or to a list of random assignments generated by a computer. Also, restricted allocation can be performed such as blocked randomization (to ensure particular allocation ratios to the intervention groups), stratified randomization (randomization performed separately within strata), or minimization (to make small groups closely similar with respect to several characteristics). Another important characteristic to judge if randomization was appropriately performed is allocation concealment that protects assignment sequence until allocation. Researchers and participants should be unaware of the assignment sequence up to the point of allocation. Several strategies can be used to ensure allocation concealment such relying on a central randomization by a third party, or the use of sequentially numbered, opaque, sealed envelopes (Higgins et al., 2016).

2.2 Are the groups comparable at baseline?

Explanations

Baseline imbalance between groups suggests that there are problems with the randomization. Indicators from baseline imbalance include: "(1) unusually large differences between intervention group sizes; (2) a substantial excess in statistically significant differences in baseline characteristics than would be expected by chance alone; (3) imbalance in key prognostic factors (or baseline measures of outcome variables) that are unlikely to be due to chance; (4) excessive similarity in baseline characteristics that is not compatible with chance; (5) surprising absence of one or more key characteristics that would be expected to be reported" (Higgins et al., 2016, p. 10).

2.3 Are there complete outcome data?

Explanations

Almost all the participants contributed to almost all measures. There is no absolute and standard cut-off value for acceptable complete outcome data. Agree among your team what is considered complete outcome data in your field and apply this uniformly across all the included studies. For instance, in the literature, acceptable complete data value ranged from 80% (Thomas et al., 2004; Zaza et al., 2000) to 95% (Higgins et al., 2016). Similarly, different acceptable withdrawal/dropouts rates have been suggested: 5% (de Vet et al., 1997; MacLehose et al., 2000), 20% (Sindhu et al., 1997; Van Tulder et al., 2003) and 30% for a follow-up of more than one year (Viswanathan and Berkman, 2012).

2.4 Are outcome assessors blinded to the intervention provided?

Explanations

Outcome assessors should be unaware of who is receiving which interventions. The assessors can be the participants if using participant reported outcome (e.g., pain), the intervention provider (e.g., clinical exam), or other persons not involved in thDid the participants adhere to the assigned intervention?e intervention (Higgins et al., 2016).

2.5 Did the participants adhere to the assigned intervention?

Explanations

To judge this criterion, consider the proportion of participants who continued with their assigned intervention throughout follow-up. "Lack of adherence includes imperfect compliance, cessation of intervention, crossovers to the comparator intervention and switches to another active intervention." (Higgins et al., 2016, p. 25).

Quantitative non-randomized studies

3.1. Are the participants representative of the target population?

Explanations

Indicators of representativeness include: clear description of the target population and of the sample (inclusion and exclusion criteria), reasons why certain eligible individuals chose not to participate, and any attempts to achieve a sample of participants that represents the target population.

3.2. Are measurements appropriate regarding both the outcome and intervention (or exposure)?

Explanations

Indicators of appropriate measurements include: the variables are clearly defined and accurately measured; the measurements are justified and appropriate for answering the research question; the measurements reflect what they are supposed to measure; validated and reliability tested measures of the intervention/exposure and outcome of interest are used, or variables are measured using 'gold standard'.

3.3. Are there complete outcome data?

Explanations

Almost all the participants contributed to almost all measures. There is no absolute and standard cut-off value for acceptable complete outcome data. Agree among your team what is considered complete outcome data in your field (and based on the targeted journal) and apply this uniformly across all the included studies. For example, in the literature, acceptable complete data value ranged from 80% (Thomas et al., 2004; Zaza et al., 2000) to 95% (Higgins et al., 2016). Similarly, different acceptable withdrawal/dropouts rates have been suggested: 5% (de Vet et al., 1997; MacLehose et al., 2000), 20% (Sindhu et al., 1997; Van Tulder et al., 2003) and 30% for follow-up of more than one year (Viswanathan and Berkman, 2012).

3.4. Are the confounders accounted for in the design and analysis?

Explanations

Confounders are factors that predict both the outcome of interest and the intervention received/exposure at baseline. They can distort the interpretation of findings and need to be considered in the design and analysis of a non-randomized study. Confounding bias is low if there is no confounding expected, or appropriate methods to control for confounders are used (such as stratification, regression, matching, standardization, and inverse probability weighting).

3.5 During the study period, is the intervention administered (or exposure occurred) as intended?

Explanations

For intervention studies, consider whether the participants were treated in a way that is consistent with the planned intervention. Since the intervention is assigned by researchers, consider whether there was a presence of contamination (e.g., the control group may be indirectly exposed to the intervention) or whether unplanned co-interventions were present in one group (Sterne et al., 2016).

For observational studies, consider whether changes occurred in the exposure status among the participants. If yes, check if these changes are likely to influence the outcome of interest, were adjusted for, or whether unplanned co-exposures were present in one group (Morgan et al., 2017).

Mixed methods studies

5.1. Is there an adequate rationale for using a mixed methods design to address the research question?

Explanations

The reasons for conducting a mixed methods study should be clearly explained. Several reasons can be invoked such as to enhance or build upon qualitative findings with quantitative results and vice versa; to provide a comprehensive and complete understanding of a phenomenon or to develop and test instruments (Bryman, 2006).

5.2. Are the different components of the study effectively integrated to answer the research question?

Explanations

Integration is a core component of mixed methods research and is defined as the 'explicit interrelating of the quantitative and qualitative component in a mixed methods study' (Plano Clark and Ivankova, 2015, p. 40). Look for information on how qualitative and quantitative phases, results, and data were integrated (Pluye et al., 2018). For instance, how data gathered by both research methods was brought together to form a complete picture (e.g., joint displays) and when integration occurred (e.g., during the data collection-analysis or/and during the interpretation of qualitative and quantitative results).

5.3. Are the outputs of the integration of qualitative and quantitative components adequately interpreted?

Explanations

This criterion is related to meta-inference, which is defined as the overall interpretations derived from integrating qualitative and quantitative findings (Teddlie and Tashakkori, 2009). Meta- inference occurs during the interpretation of the findings from the integration of the qualitative and quantitative components, and shows the added value of conducting a mixed methods study rather than having two separate studies.

5.4. Are divergences and inconsistencies between quantitative and qualitative results adequately addressed?

Explanations

When integrating the findings from the qualitative and quantitative components, divergences and inconsistencies (also called conflicts, contradictions, discordances, discrepancies, and dissonances) can be found. It is not sufficient to only report the divergences; they need to be explained. Different strategies to address the divergences have been suggested such as reconciliation, initiation, bracketing and exclusion (Pluye et al., 2009b). Rate this criterion 'Yes' if there is no divergence.

5.5. Do the different components of the study adhere to the quality criteria of each tradition of the methods involved?

Explanations

The quality of the qualitative and quantitative components should be individually appraised to ensure that no important threats to trustworthiness are present. To appraise 5.5, use criteria for the qualitative component (1.1 to 1.5), and the appropriate criteria for the quantitative component (2.1 to 2.5, or 3.1 to 3.5, or 4.1 to 4.5). The quality of both components should be high for the mixed methods study to be considered of good quality. The premise is that the overall quality of a mixed methods study cannot exceed the quality of its weakest component. For example, if the quantitative component is rated high quality and the qualitative component is rated low quality, the overall rating for this criterion will be of low quality.

Appendix E Reporting the results of the MMAT (version 2018)

In the version 2018, we advised not to present an overall score. This decision was made from the literature that discouraged to use metrics because it is not informative. By presenting a single number, it is not possible to know what aspects of studies are problematic. We often see people presenting a global score and nothing else in the results or discussion or description of included studies. This often raises the question of why quality appraisal was performed.

This suggestion is, however, problematic for reporting the results of the MMAT. Several MMAT users have contacted us for advice to report their results. If there is a need to report an overall score, here is a suggestion based on the previous version of the MMAT:

For each retained study, an overall quality score may not be informative (in comparison to a descriptive summary using MMAT criteria), but might be calculated using the MMAT. Since there are only a few criteria for each domain, the score can be presented using descriptors such as stars (*) or %:

5***** or 100% quality criteria met

4 **** or 80% quality criteria met

3 *** or 60% quality criteria met

2 ** or 40% quality criteria met

1 * or 20% quality criteria met

For mixed methods studies, since there are 15 criteria to rate (instead of 5), the premise is that the overall quality of a combination cannot exceed the quality of its weakest component. Thus, the overall quality score is the lowest score of the study components.

The score is 20% (*) when QUAL=1 or QUAN=1 or MM=1;

it is 40% (**) when QUAL=2 or QUAN=2 or MM=2;

it is 60% (***) when QUAL=3 or QUAN=3 or MM=3;

it is 80% (****) when QUAL=4 and QUAN=4 and MM=4,

and it is 100% (*****) when QUAL=5 or QUAN=5 or MM=5; (QUAL being the score of the qualitative component; QUAN the score of the quantitative component; and MM the score of the mixed methods component).

Regarding questions on cut off value, we have not studied values that could characterize low, medium or high quality studies. The categories are arbitrary, but useful for performing qualitative or quantitative sensitivity analysis. We have seen some papers with 2 categories (lower vs higher quality) or 3 categories (e.g., low, medium, and high). What is important is to clearly describe how the results of the appraisal were interpreted and used in the review (transparency).

Appendix F List of references of included studies

1. Aclan, E. M., & Aziz, N. H. A. (2015a). Exploring parliamentary debate as a pedagogical tool to develop English communication skills in EFL/ESL classrooms. *International Journal of Applied Linguistics and English Literature*, *4*(2), 1-16.

2. Al-Tamimi, N. (2017). Improving speaking skill: Implications for using debatable topics in English speaking classes. *English Language Teaching and Literature. Retrieved from https://www. researchgate. net/publication/320853097*.

3. Bahrani, B., Syatriana, E., Afdaliah, N., Syamsinar, S., & Supriadi, B. (2020). A method to push students' speaking ability in English foreign classroom. *The Asian ESP Journal, 16*(62), 5-12.

4. Baso, F. A. (2016). The implementation of debate technique to improve students' ability in speaking. *Exposure, 5*(2), 154-173.

5. Brown, N. A., Talalakina, E. V., Yakusheva, I. V., & Eggett, D. L. (2012). Argumentation and debate in the foreign language classroom: Russian and American university students collaborating through new technologies. *Russian Language Journal, 62*, 141-166.

6. Dehham, S. H. (2021). Enhancing Iraqi EFL preparatory school students' speaking skill by using debate technique. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(7),3189-3196.

7. Desita, L. R., Supardi, I., & Suhartono, L. (2017). Improving students' speaking ability through debate technique. *Jurnal Pendidikan dan Pembelajaran Khatulistiwa (JPPK), 6*(5).

8. El Majidi A., de Graaff R., Janssen D. (2018). Students' perceived effect of in-class debates in second language learning. *The European Journal of Applied Linguistics and TEFL, 7*, 35–57.

9. El Majidi A., de Graaff R., Janssen D. (2021a). Debate as a pedagogical tool for developing speaking skills in second language education. *Language Teaching Research*. Epub ahead of print 10 October 2021.

10. El Majidi A., de Graaff R., Janssen D. (2021b). The effects of in-class debates on argumentation skills in second language education. *System, 101*, 102576.

11. Elmiyati, E. (2019). Improving students speaking ability through debate in the classroom (A case study for students at second years students of SMAN3 Kota Bima in Academic Year 2017/2018). *International Journal of Multicultural and Multireligious Understanding, 6*(1), 229-239.

12. Fauzan, U. (2016). Enhancing speaking ability of EFL students through debate and peer assessment. *EFL journal, 1*(1), 49-57.

13. García-Sánchez, S. (2020). Debates in English language education: A multimodal, collaborative ecosystem. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT), 10*(4), 33-50.

14. García-Sánchez, S., & Gimeno-Sanz, A. (2022). Telecollaborative debates in ESP: Learner perceptions and pedagogical implications. *Journal of Teaching English for Specific and Academic Purposes*, 139-157.

15. He, J. (2015). The effects of an argumentation and debate course on students' impromptu English speaking competence: An empirical investigation. *English Teaching and Learning, 39*(3), 79-106. DOI: 10.6330/ETL.2015.39.3.04

16. Iman, J. N. (2017). Debate instruction in EFL classroom: Impacts on the critical thinking and speaking skill. *International Journal of Instruction, 10*(4), 87-108.

17. Kassem, H. M. (2021). Training EFL learners on debating: Effects on their oral and written. *MEXTESOL Journal, 45*(4).

18. Khalil, E. F. A. (2018). *The effects of debate instruction on Turkish EFL learners' L2 speaking anxiety, L2 speaking performance, and L2 writing performance* (Master's thesis). Bahçeşehir Üniversitesi, İstanbul.

19. Lustigova, L. (2011). Speak your mind: Simplified debates as a learning tool at the university level. *Journal on Efficiency and Responsibility in Education and Science, 4*(1), 18-30.

20. Nobuhara, M. (2016). Analysis of debate classes and potential factors related to leading Japanese universities' rducation. *BESSH-2016 Full Paper Proceeding, 114*(2), 49-53. Retrieved from http://academicfora.com/wp-content/uploads/2016/04/FULL-PAPER-PROCEEDING-TKS-436-124-1.pdf

21. Noor, S., & Tajik, O. (2022). Empowering speaking skill through debate: The case of Afghan EFL learners at Herat university. *International Journal of Multidisciplinary: Applied Business and Education Research, 3*(12), 2631-2640.

22. Putri, A. A., & Rodliyah, R. S. (2021). EFL Students' perception on the use of debate in speaking classroom. *The Thirteenth Conference on Applied Linguistics (CONAPLIN 2020), 546* (Conaplin 2020), 21–29.

23. Rosas-Maldonado, M., Bascuñán, D., & Martin, A. (2020). Potential benefits of debates: Perceptions of EFL pre-service teachers. *Lenguas Modernas,* (55).

24. Sabbah, S. R. M. (2015). *The effectiveness of using debates in developing speaking skills among English majors at University of Palestine* (Master's thesis). Al-Azhar University – Gaza.

25. Shamsudin, M., Othman, M., Jahedi, M., & Aralas, D. (2017). Enhancing English learners' willingness to communicate through debate and philosophy inquiry discussion. *English Language Teaching, 10*(8), 145-152.

26. Shan, L. (2005). *The effect of debate on oral communication skills among university students in Taiwan: A case study* (Master's thesis). National Tsing Hue University.

27. Tianame, M., Usman, B., & Muslem, A. (2019). The implementation of debate technique to enhance the students' speaking skill. *English Education Journal, 10*(2), 180-195.

28. Yulia, H., & Aprilita, N. (2017). The implementation of British parliamentary debate style training to improve second semester student's speaking ability at English education study program of Baturaja University. *International Journal of Language Teaching and Education, 1*(1), 1-7.

29. Zare, P., & Othman, M. (2015). Students' perceptions toward using classroom debate to develop critical thinking and oral communication ability. *Asian Social Science, 11*(9), 158.

Appendix G Basic information of included studies

Study	Publication type	Location	Education phase	Research methodology and Data type	Sample size	Intervention duration	Debate format and control group (if any)	Methods of data collection	Reported outcomes
1. Aclan & Aziz (2015a)	Journal article	Association of Southeast Asian Nations (Indonesia, Malaysia, Myanmar, Philippines, and Thailand)	Tertiary	- Case study - Qualitative	11: Experts (n=6) Students (n=5)	- Experts: at least two years - Students: 16 weeks, four times a week at 1.5 hours per day	Asians Parliamentary Debate format	Semi-structured one-on-one interview and focus group interview	- Language aspect: coherence, fluency, accuracy, vocabulary, speech length - Affective aspect: motivation, confidence - Challenges: Point of information (POI) and rebuttals
2. Al-Tamimi (2017)	Journal article	Yemen	Tertiary	- The pre-test-post-test control group quasi-experimental design - Quantitative	70: Intervention (n=35) Control (n=35)	16 weeks	- Not specified - Control: regular speaking instruction	- Speaking tests (picture description, question and answer, topic discussion) - Questionnaire	- Language aspect: pronunciation, accuracy, fluency, vocabulary, interaction and involvement - Affective aspect: enjoyment, motivation
3. Bahrani et al. (2020)	Journal article	Indonesia	Secondary	- The pre-test-post-test control group quasi-experimental - Quantitative	50: Intervention (n=26) Control (n= 24)	12 weeks	- Researcher-designed debate pedagogy - Control: the conventional method	Interview sheet	- Language aspect: general speaking proficiency

4. Baso (2016)	Journal article	Indonesia	Secondary	- The one-group pre-test-post-test pre-experimental design - Quantitative	20	Not specified	Four-corner debate format	Responsive speaking tests	- Language aspect: accuracy, fluency
5. Brown (2012)	Journal article	Russia and the US	Tertiary	- The pre-test-post-test quasi-experimental design with two groups - Quantitative	30: Russian EFL students (n=12) American RFL students (n =18)	One semester	Video-conference debate	Oral proficiency interview	- Language aspect: argumentation (opinion expression and support; speculation and hypotheses), interaction and involvement (engage in extended discussions)
6. Dehham (2021)	Journal article	Iraq	Secondary	The pre-test-post-test control and experimental group design - Quantitaitve	60: Intervention (n=30) Control (n=30) All female	Nine weeks	- Not specified - Control: the regular way using textbook	Speaking tests	- Language aspect: general speaking proficiency
7. Desita et al. (2017)	Master's thesis	Indonesia	Secondary	The one-group pre-test-post-test pre-experimental design - Quantitative	26	One semester, three cycles	Not specified	- Observation checklist - Field notes	- Language aspect: argumentation (giving opinion and responding to others' opinion)
8. El Majidi et al. (2018)	Journal article	The Netherlands	Secondary	- Case study - Qualitative and Quantitative	134	10 weeks, 50 minutes per week	Four debate formats with different team sizes, speech duration, and questioning periods	- Questionnaire - Semi-structured one-on-one and group interviews - Oral performance assessment	- Language aspect: argumentation (constructive speech; reasoning; rebuttal/crossfire), fluency, vocabulary, grammar, interaction and involvement (preparation) - Affective aspect: enjoyment, confidence, reduced anxiety

9. El Majidi et al. (2021a)	Journal article	The Netherlands	Secondary	- The pre-test-post-test control group quasi-experimental design - Quantitative	147: Intervention (*n*=96) Control (*n*=51)	10 weeks, 10 debates, one per week	- Researcher-designed debate pedagogy - Control: face-to-face discussions, but more reading and writing	Speaking tests: argumentative opinion tasks	- Language aspect: speech length, fluency, syntactic complexity, lexical complexity, accuracy, cohesion
10. El Majidi et al. (2021b)	Journal article	The Netherlands	Secondary	- The pre-test-post-test control group quasi-experimental design - Quantitative	147: Intervention (*n*=96) Control (*n*=51)	10 weeks, 10 debates, one per week	-Researcher-designed: one-on-one and two-on-two debate format - Control: face-to-face discussions, but more reading and writing	Speaking tests: argumentative opinion tasks	- Language aspect: argumentation (structural analysis; reasoning quality)
11. Elmiyati (2019)	Journal article	Indonesia	Secondary	- The pre-test-post-test control group experimental design - Quantitative	60: Intervention (n=30) Control (n=30)	Not specified	- Not specified - Control: no treatment	- Speaking tests - Questionnaire	- Language aspect: pronunciation/accent, grammar, vocabulary, fluency, comprehension
12. Fauzan (2016)	Journal article	Indonesia	Tertiary	- The one-group pre-test-post-test pre-experimental design - Qualitative and Quantitative	Not specified	Two cycles, eight meetings, four meetings in each cycle	British Parliamentary debate format	- Observation checklist - Field-note, - Score sheet - Questionnaire	- Language aspect: fluency, speech length - Affective aspect: confidence
13. García-Sánchez (2020)	Journal article	Spain	Tertiary	- Case study - Qualitative and Quantitative	126	15 weeks	Researcher-designed debate pedagogy (computer-assisted)	- Online schemes/scripts written by the teams - Observation of mock debates - Recorded videos in the final debate	- Language aspect: argumentation (organization and clarity; use of arguments; use of examples and facts; use of rebuttals), delivery (tone of voice, use of gestures, level of enthusiasm) - Affective aspect: confidence

								- Questionnaire	
14. García-Sánchez & Gimeno-Sanz (2022)	Journal article	Spain	Tertiary	- The one-group pre-test-post-test pre-experimental design - Qualitative and Quantitative	25	15 weeks, 90 minutes per week	Video-conference debate	- Pre- and post-project anonymous survey - Observation of face-to-face and online group discussions	- Language aspect: general speaking proficiency, vocabulary, grammar - Affective aspect: confidence, motivation
15. He (2015)	Journal article	China	Tertiary	- The pre-test-post-test control group quasi-experimental design - Quantitative	53: Intervention (n=30) Control (n=23)	18 weeks, 90 minutes per week	- British Parliamentary Debate format - Control: group discussions and presentations	Speaking tests: three-minute impromptu speeches	- Language aspect: vocabulary (lexical diversity; lexical richness), syntactic complexity
16. Iman (2017)	Journal article	Indonesia	Secondary	- The pre-test-post-test control group quasi-experimental design - Quantitative	48: Intervention (n=24) Control (n= 24) All female	One academic year with 21 teaching sessions, 90 minutes each session	- World School Debate Championship format - Control: no treatment	Speaking tests: monologue measured via Student Oral Language Observation Matrix	- Language aspect: comprehension, vocabulary, pronunciation, grammar, fluency
17. Kassem (2021)	Journal article	Saudi Arabia	Tertiary	- The one-group pre-test-post-test pre-experimental design - Quantitative	32 male	Eight weeks	Researcher-designed debate pedagogy	Speaking tests: opinion task	- Language aspect: vocabulary, cohesion, fluency, argumentation (relevance and adequacy of content; adequacy and strength of arguments) - Affective aspect: motivation

									(ideal L2 self), anxiety (communication apprehension)
18. Khalil (2018)	Master's thesis	Turkey	Secondary	- The one-group pre-test-post-test pre-experimental design - Qualitative and Quantitative	50: A1 (n=26) A2 (n=24)	Eight weeks	Researcher-designed debate pedagogy	- Speaking tests: picture descriptions - Questionnaire - Filed notes -Semi-structured interviews	- Language aspect: clarity, argumentation, vocabulary - Affective aspect: confidence, anxiety - Challenges: motivation, satisfaction and learners' proficiency
19. Lustigova (2011)	Journal article	The Czech Republic	Tertiary	Case study - Qualitative and Quantitative	475 Class sizes varied, from 20 to 25 students	Two semesters	Researcher-designed: simplified debate formats	Observation	- Language aspect: general speaking proficiency, speech length - Affective aspect: confidence, anxiety
20. Nobuhara (2016)	Conference proceedings	Japan	Tertiary	- The one-group pre-test-post-test pre-experimental design - Qualitative and Quantitative	38	Half of an academic year	Researcher-designed debate pedagogy	- Questionnaire - Interview	- Language aspect: general speaking proficiency, fluency - Affective aspect: motivation
21. Noor & Tajik (2022)	Journal article	Afghanistan	Tertiary	- The one-group pre-test-post-test pre-experimental design - Quantitative	44	Six weeks	Not specified	- Speaking tests - Questionnaire	- Language aspect: general speaking proficiency (general description; language use; topic development), delivery, comprehensibility, argumentation (topic development) - Affective aspect: anxiety, enjoyment

22. Putri & Rodliyah (2021)	Conference proceedings	Indonesia	Tertiary	- Case study - Qualitative	16: Questionnaire (n = 10) Interview (n = 6)	One semester	Not specified	- Questionnaire - Interview	- Language aspect: fluency - Affective aspect: confidence, anxiety
23. Rosas-Maldonado et al. (2020)	Journal article	Chile	Tertiary	- Case study - Qualitative	17	Five weeks, two hours in each class	Parliamentary debate format	Focus group interview	- Language aspect: fluency, grammatical accuracy, vocabulary - Affective aspect: motivation, enjoyment, confidence
24. Sabbah (2015)	Master's thesis	Palestine	Tertiary	- The one-group pre-test-post-test pre-experimental design - Qualitative and Quantitative	20	One semester, 17 sessions, 30 minutes in each session	Researcher-designed debate pedagogy	- Speaking tests: opinion tasks - A teacher's guide	- Language aspect: pronunciation, grammar, vocabulary
25. Shamsudin et al. (2017)	Journal article	Malaysia	Tertiary	- The pre-test-post-test control group experimental design - Quantitative	32: Intervention (n=16) Control (n=16)	Eight weeks	- Researcher-designed debate pedagogy - Control: philosophy inquiry discussion	Questionnaire	- Affective aspect: motivation (willingness to communicate), confidence (self-communicative competence), anxiety (communicative apprehension)
26. Shan (2005)	Master's thesis	China	Tertiary	- The pre-test-post-test control group quasi-experimental design - Qualitative and Quantitative	21: Intervention (n=11 or 10) Control (n=10 or 11) All female	Six weeks debate for intervention group, six weeks discussion for control group, then exchanged	- Cross-examination format - Control: discussion, presentations, film watching, discussion, charades, role plays	- Speaking tests: three minutes impromptu speeches -Questionnaires - Student journals - Interview	- Language aspect: argumentation (speech development; effectiveness of communication), delivery (physical aspects; voice), general speaking proficiency (language use) - Affective aspect: confidence, anxiety

27. Tianame et al. (2019)	Journal article	Indonesia	Secondary	- The one-group pre-test-post-test pre-experimental design - Qualitative and Quantitative	27	Four weeks, two cycles with four meetings	Researcher-designed debate pedagogy	- Speaking test - Observation checklist - Questionnaire	- Language aspect: general speaking proficiency, involvement
28. Yulia & Aprilita (2017)	Journal article	Indonesia	Tertiary	- The one-group pre-test-post-test pre-experimental design - Quantitative	12	Not specified	British Parliamentary debate format	Not specified	- Language aspect: fluency, grammar, pronunciation, vocabulary, comprehension
29. Zare & Othman (2015)	Journal article	Malaysia	Tertiary	- Case study - Qualitative and Quantitative	16	One semester, 9 sessions, 60 minutes per session	British Parliamentary debate format	- Questionnaire - Reflective papers - Interviews	- Language aspect: general speaking proficiency - Affective aspect: confidence, anxiety - Challenges: topic selection and preparation time

Appendix H Sample characteristics

Study ID	Education phase	English proficiency	Sampling strategies
1	Tertiary	Average TOEFL score 391 (approximately A2 to B1 level)	Purposive sampling: participants were selected in to the debate class based on high scores in listening and speaking entrance tests
2	Tertiary	Third-year English major students L1: Native speakers of Arabic	Convenience sampling: two intact classes
3	Secondary		Convenience sampling: two intact classes
4	Secondary		Purposive sampling
5	Tertiary	All Russian EFL participant had already passed an English language entrance exam and completed 72 hours of ESL-related courses during their first year of study. They were also enrolled in *Business English* alongside the debate course	
6	Secondary		Randomization
7	Secondary	B1 to B2	Convenience sampling: seven intact classes
8	Secondary	B1	Convenience sampling: eight intact classes
10	Secondary	B1 to B2	Convenience sampling: eight intact classes
11	Secondary		Randomization

12	Tertiary	Third-semester students at the English Department	
13	Tertiary	- B1+ to B2 - Debate pedagogy in EFL compulsory subject: *Communicative Competences in English* - fourth-year degree in *Telecommunications Engineering*	
14	Tertiary	- B1+ to B2 - Debate pedagogy in *English for Specific Purposes (ESP)* course including *English for Telecommunications* (n = 7) and *Specialised English* (n = 18) - Second year students of Master's Degree in *Telecommunications Engineering* (n = 7); Third-year students in *Aerospace Engineering Degree* (n = 18)	
15	Tertiary	- Debate pedagogy in *Argumentation and Debate* open for Master's students from all departments (e.g., *Environmental Science, Life Science, Philosophy, Chemistry, Information Technology, Literature*) in a research university	Voluntary basis for debate course
16	Secondary		Purposive sampling
17	Tertiary	Participants have taken three speaking courses but not debate training prior to the intervention	Convenience sampling: an intact class
18	Secondary	- A1 and A2 - Participants have been learning	Voluntary basis for debate course

		English from 7 to 10 years	
19	Tertiary	- Debate pedagogy in *General English* at the University intermediate and lower-level classes, a required course for B.A. degree course - Major varied (e.g., *Economics and Management*, *Engineering*, *Environmental Science*, *Forestry*, *Wildlife and Wood Sciences*, *Agrobiology*, *Food and Natural Resources*) - L1: native speakers of Czech, but some of them also came from multinational backgrounds (including Russian, Ukrainian, Kazakh and Vietnamese families)	
20	Tertiary	- Leading Japanese university ESL students	Convenience sampling
21	Tertiary	- English department faculty - L1: mostly Dari and Pashto languages	
22	Tertiary	- Debate pedagogy in *Speaking for Academic Purposes Course*	Convenience sampling
23	Tertiary	- B1 - Second Language Teacher Education programme - L1: Spanish	Voluntary basis for debate course
24	Tertiary	English Major students	Purposive sampling
25	Tertiary	Debate pedagogy in *Oral and Aural*	Randomization

		Course	
26	Tertiary	- Participants had been learning English for at least eight years - *English* majors (n = 20) and *Electrical Engineering* (n = 1) students from an advanced oral communication course	
27	Secondary		Convenience sampling: an intact class
28	Tertiary	English Education Study Program	
29	Tertiary	Teaching English as a Second Language program	

Appendix I Intervention details

Study ID	Debate format	Pre-debate activity	Actual debate	Post-debate activity	Control group (if any)
1	Asians Parliamentary Debate format	- Brainstorming with team member - Read relevant literature - Outline their speech, write note	- Debaters: present arguments, challenge and engage in point of information (POI) and rebuttals - Audience: present POI	- Adjudicators/teachers: give oral corrective feedback and comments - Debaters: interact with adjudicators, team-mates and opponents	/
2	Not specified	Not specified	Not specified	Not specified	Regular speaking instruction, but not specified
3	Researcher-designed debate pedagogy	- Researchers teach debate techniques such as how to think critically, how to give a good statement, argument and solution towards every motion that want to solve	- Debaters: present arguments - Researcher: observe	Not specified	The conventional method, but not specified
4	Four-corner debate format (Four corners of the classroom are	- Teachers present a question or statement	- Students within each corner collaborate to develop arguments to support their	Not specified	/

	labelled as strongly agree, agree, disagree, and strongly disagree. Students move to the corner that aligns with their position on the issue)	- Students individually consider the statement and form their views	position - Group discussions take place within each corner to refine arguments - Each group presents their strongest arguments/rebuttals to the other groups - Students have the option to switch sides if their personal views have changed		
5	Video-conference debate	- Students' learn principles of argumentation and debate in the students' first language - Researchers distribute ACTFL oral and written proficiency guidelines to participants - Six Russian and six U.S. students are assigned to each debate	- U.S. and Russian students debate via synchronous video-conference technology - Language alternates each week, allowing both sides to use Russian and English - Each speaker have four minutes to present their position on a given topic. - Approximately 50 minutes per round of debate.	- Asynchronous written blog discussions carried out in the language of the oral debate - Participants have the option to opt in or out of blog discussions - Assigned weekly homework included writing and reading assignments	/
6	Not specified	Not specified	Not specified	Not specified	The regular way using textbook

7	Not specified	Not specified	Not specified	Not specified	/
8	Four debate formats with different team sizes, speech duration, and questioning periods.	- Students are given one week to prepare for each debate. - Read and find articles on the topic - Submit a debate preparation form that include their constructive speech, arguments for both sides, and sources used	- Debate Format 1: Two teams with affirmative and negative sides. Each student has three cards, handing in one each time they speak - Debate Format 2: Two teams with two speakers each. They deliver one-minute constructive speeches, followed by a two-minute rebuttal and a crossfire questioning period. - Debate Format 3: Two debaters, one on the affirmative side and the other on the negative side. Students have one minute time in the first two rounds, three minutes in the third round and five minutes' preparation time between the constructive speech and rebuttal. - Debate Format 4: No outside preparation is required for this format. Topics are determined randomly, pairs are formed. Students have five minutes to	Not specified	/

			prepare and make notes for the debate. Each debater has the chance to present their arguments, counter arguments, and cross-examine their opponent. The first two rounds allow for 30 seconds of speaking time, while the third round provides three minutes. There is a five-minute preparation period between the constructive speech and rebuttal.		
9	Researcher-designed debate pedagogy	- Topics for debates are chosen in consultation with the debating students - Reading assignment - Summarise the articles and write a case defending their viewpoints - One week of preparation time for each debate	- Debates consists of three phases: speech, rebuttal, and clash - Each student presents arguments and rebuttals - Students participate in a clash, listen to each other and take note of mistakes and new words learned	- Teachers provide feedback on the students' written cases - Students revise and resubmit their cases for further improvement	- Engage in traditional controlled face-to-face discussions - Regular four language skills instruction with more reading and writing assignments
10	Researcher-designed: One-on-one debate format, two-to-two	- Students are informed at least one week in advance of the debate	- Students present constructive speech, rebuttal and clash	Not specified	- Engage in traditional controlled face-to-

	debate format (a group of four debaters, two students in favor and two against)	topic - Debate topics and the side to defend (affirmative or negative side) are based on students' own choices - Read two articles (related to the topic under debate) and summarise them. The instructor provided one article and the students had to find another one - Write persuasive essays and letters	- Students note down the arguments of the opponents		face discussions - Regular four language skills instruction with more reading and writing assignments
11	Not specified	Not specified	Not specified	Not specified	No treatment
12	British Parliamentary debate format	- Researcher plan the implementation of debate and peer assessment, prepare the teaching strategy, design a lesson plan, provide media, make an assessment sheet, and	- Cycle 1 - Debaters deliver 7 minutes speeches in the following order: (1) Prime Minister, (2) Leader of the Opposition, (3) Deputy Prime Minister, (4) Deputy Leader of the Opposition, (5) Government Member, (6) Opposition Member, (7) Government	- Researchers evaluate the result of the implementation of the strategy and draw a conclusion on whether the strategy is successful or not by comparing it to the criteria of success	/

		determine the criteria of success - Four teams in each debate round, and each team has two members.	Whip, and (8) Opposition Whip - Students are to assign assessments and give feedback to their peers. - Each group are assessed by two students Cycle 2 - Choose the simpler motion - Instruct the students to find the data related to the motion - Give more time to rehearse in pairs - Let the students write the outline of their arguments - Each student assesses one group only		
13	Researcher-designed debate pedagogy (computer-assisted)	- Four sessions for debate technique training - Face-to-face sessions to exchange information accessed through CALL/MALL and plan their interventions	Debaters present arguments	Self-awareness and evaluation of skills	/

		- Optional script creation using Google Docs for collaborative work and feedback from the teacher - Individual research, writing and oral practice - Individual and team writing and oral practice for debate			
14	Video-conference debate	- Two synchronous 20-minute online sessions for understanding telecollaborative tasks, debating skills, and team work - Individual/team research, and ESP vocabulary - Individual writing and oral practice for debate - Team planning, scriptwriting, and	- Two separate break-out rooms via video-conference calls - The moderator introduces topic, teams and subtopic - Debaters present arguments: - Arguments in favor (2 mins) - Arguments against (2 mins) - Arguments in favor (2 mins) - Arguments against (2 mins) - Break to prepare rebuttals (5 mins) - Rebuttal: three persons in each team	Reflection of debate based on recorded video	/

		corrections - Digital tools: WhatsApp, Skype, Google Hangouts for synchronous communication, Google+, Google Forms for asynchronous communication - Team roles: moderator, team in favor (10 members), team against (10 members), audience (4 members)	give rebuttals with no time limits - Question: The audience poses two questions to each team - The moderator makes concluding remarks and closes session		
15	British Parliamentary Debate format	pre-test: three-minute impromptu speeches	Not specified	Post-test: three-minute impromptu speeches	Traditional speaking method using group discussions and presentations
16	World School Debate Championship format (two teams (affirmative/negative), each consists of	- The teacher introduces debate format and technique - The teacher selects debaters, chairperson,	- Chairperson opens the debate - Each speaker presents speech and rebuttals in the following order: - Substantive speech, 8 minutes each: (1)	- The teacher asks the student to summarize the debate - The teacher adds the summary from their debate	No treatment

	three debaters)	time keeper, odd number of judges - Debate motion is released at the end of the former debate meeting	1st speaker in affirmative team, (2) 1st speaker in negative team, (3) 2nd affirmative speaker, (4) 2nd negative speaker, (5) 3rd affirmative speaker, (6) 3rd negative speaker - Reply speech, 4 minutes each: (7) 1st or 2nd affirmative speaker; (8) 1st or 2nd negative speaker - POI: During the substantive speech, members of the opposing team can give POI, no POIs are allowed in a reply speech, the speaker can accept or reject a POI - Chairperson close the debate	- The teacher informs the debate motion for the next debate meeting	
17	Researcher-designed debate pedagogy	- Eight groups of four students each, with one student being the group's coordinator - Each group is assigned as the affirmative or the negative side based on their own choice - Every group do eight	- The affirmative position presents their first argument - The negative position rebut with counter-arguments and present their arguments - The procedure is repeated for all arguments and counter-arguments - Each group's coordinator sum up the	- The researcher teach how to organize an argumentative essay - Write an essay about the issue and submit it within three days - Feedback given by the researchers	/

		debates and write eight argumentative essays in sessions where they were no debaters - Students are shown videotaped debates to get an idea about the notion and procedure of debating - Researchers teach debate techniques and research skills - Students research literature	arguments - Audience voice their opinion about which group is more convincing		
18	Researcher-designed debate pedagogy	- Students are divided randomly into two groups (affirmative or negative side), 3 to 4 members in each team one week before the next debate session - Read relevant literature - Warm up questions (5 minutes)	- 1 st speaker in affirmative team: introduce the debate topic and defined their position - 1st negative introduce topic and position - 2nd affirmative: rebuttal to Negative, present arguments - 2nd negative: rebuttal to Affirmative side, present arguments - 3rd affirmative: rebuttal to Negative,	The teacher trains students to write about different debatable topics (10 minutes)	

		- Learn new vocabulary by using pictures (3 minutes)	present arguments - 3rd negative: rebuttal to Affirmative side, present arguments - Last affirmative and negative: conclude each side - Approximately 40 minutes		
19	Researcher-designed: Simplified debate formats	- Two teams of two or three members, team's role are either assigned by the teacher or the students' own choice - Pre-reading - Topic definition - Warm-up discussion - Comprehension exercises - Read, watch and listen to recorded debates or videos on a given topic - Vocabulary and phrases review	- Affirmative team speech - Negative team speech - Affirmative team rebuttal - Negative team rebuttal - Questions/answers from teams, field questions from audience - Affirmative closing argument - Negative closing argument	- Audience assesses arguments' persuasiveness - Teacher provides constructive feedback - Homework for writing and literature research	/

20	Researcher-designed debate pedagogy	- 4 to 6 students per group - Ten minutes for preparation	- Affirmative and negative sides presents arguments - Approximately 35 minutes	- Self-reflection - Teachers feedback	
21	Not specified	Not specified	Not specified	Not specified	/
22	Not specified	Not specified	Not specified	Not specified	/
23	Parliamentary debate format	- The teacher introduced debate format and technique - Language practice - Three groups of 4 students, one group of five students	- Two teams debating while the other groups act as audience and evaluators of debaters performance, ask questions to debate teams	Not specified	/
24	Researcher-designed debate pedagogy	- Warm-up revision: revise some structures about expressing opinion, agreeing and disagreeing - The teacher introduces debate terms and procedure, motion, and	- Debaters present arguments, rebuttals - Summarize the debate	The teacher and students assesses strengths and weaknesses of the debate	/

		give instructions - Two teams (affirmative or negative), two person in a team			
25	Researcher-designed debate pedagogy	- The teacher introduces debate format, role of each speaker,debate technique - Students read literature on the topics that would be discussed in the next debate sessions - 30 minutes brainstorm and discuss their points	Debaters present arguments	The teacher give an evaluation of students' delivery and argument	Philosophy inquiry discussion: - Students seat in circle and are each given a text - They read out loud paragraph by paragraph - They raise controversial issues pertaining to the text - Students are given the text beforehand so that they come prepared
26	Cross-examination Format (members of the opposing team	- Students learn about debate terms, formats, criteria, note-taking	- Speaking order and time allocation (25 mintues per round):	- Discussion of strengths and weaknesses of arguments (30	Discussion, presentations, movie watching,

	have the opportunity to ask questions to the speaker who just delivered their speech)	skills - Brainstorming sessions are conducted to decide on a debate topic - Students are randomly divided into affirmative teams and negative teams - Two-person in one debate team with others participating as timekeepers, moderators, or judges	Preparation (3 minutes) - 1st affirmative speaker presents constructive speech (3 minutes) - Cross-examination by 2nd negative (1 minute) - 1st negative speaker presents constructive speech (3 minutes) - Cross-examination by 1st affirmative 1 min - 1 min preparation time - 2nd affirmative presents constructive speech (3 minutes) - Cross-examination by 1st negative (1 minute) - 1 min preparation time - 2nd negative presents constructive speech (3 minutes) - Cross-examination by 2nd affirmative 1 minute - 1 min preparation time	minutes) - Anonymous voting for best debater, best team and best judge - Topic and role assignments are decided for the following week's debate session - Homework include watching recorded debates, practising note-taking, writing journals, providing comments, finding argumentative essays for expressions of agreement and disagreement, finding articles related to the topic	discussion, charades, role plays

			- 1st negative speaker presents rebuttal-defense-summary speech (2 minutes) - 1 min preparation time - 1st affirmative speaker presents rebuttal-defense-summary speech (2 minutes) - Timekeeper reports (1 minute) - Judges' comments (5 minutes) - Closing (1 minute)		
27	Researcher-designed debate pedagogy	Researchers teach debate technique	Students present arguments	Reflection and evaluation of the strength and weakness	/
28	British Parliamentary debate format	- Brainstorming session on the topic, exploring various interpretations and arguments - If possible, listen to each team debate another, taking opposing sides of the resolution and discussing the strong and weak points	- Each speaker delivers a speech of seven minutes and twenty seconds	- The teacher provide feedback on style, rules, delivery, and structure of speeches - Ask debaters about comments receive from judges	/

		- Four teams: Opening Government, Closing Government, Opening Opposition, and Closing Opposition - Each team has two speakers			
29	British Parliamentary debate format	- The teacher introduces debate format, rules, and debate topics - Students have one week preparation time to research both sides of the debate topic and gather relevant evidence and data - Two teams (Government and Opposition) randomly assign to the participants, with Government supports the resolution, and Opposition opposes it.	Debaters deliver a speech in the following order: (1) Prime Minister, (2) Leader of the Opposition, (3) Deputy Prime Minister, (4) Deputy Leader of the Opposition, (5) Government Member, (6) Opposition Member, (7) Government Whip, and (8) Opposition Whip - POIs are offered by opposing team members during speeches, which can be accepted or denied by the speaker. - Each debate lasting about one hour	Not specified	/

		- They had 20 to 30 minutes for team organization and preparation.			

Appendix J Discussed topics

Study	Motion	Topic words
4. Baso (2016)	Constitutional health law	Law
5. Brown (2012)	(1) Big vs. Small government, (2) Individual freedom vs. Public control, (3) Two-party vs. Multi-party (4) System secularism vs. theocracy, (5) Social Darwinism vs. Affirmative action, (6) Flat vs. Progressive tax (7) Gun control vs. Gun rights, (8) Universal vs. Private healthcare	Government, freedom, party, religion, equity, tax, gun, healthcare
9. El Majidi et al. (2021a)	The right to bear arms; Smoking should be banned	Arms, smoking
10. El Majidi et al. (2021b)	School uniforms should be introduced in schools; Smoking should be banned	Uniforms, smoking
12. Fauzan (2016)	- Cycle 1 (1) Indonesia needs a woman leader, (2) Husbands and wives should have the same educational level - Cycle 2 (3)Smoking should be permitted in public places, (4) Government should pay for educational fees, (5) Love is more important than money	Women leader, gender equality, education, smoking, education fee, love, money
13. García-Sánchez (2020)	(1) The use of technologies for EFL acquisition, (2) Should governments spy people in order to prevent terrorist attacks? (3) Impact of new technologies in society, (4) Nuclear energy, is it the future of energy suppliers? (5)Technology, the end of human work?	Technology, terrorist attacks, technology, energy, technology, work

14. García-Sánchez & Gimeno-Sanz (2022)	E.g., How will technologies change society: Is space garbage justified for connecting the world/for communications?	Technology, space garbage
15. He, J. (2015)	Topics for impromptus speech: (1) An impressive modern technology, (2) Father's love, (3) Gaokao (China's College Entrance Examination) should be abolished, (4) What I have learned from failure, (5) Being a full-time housewife is a social progress, (6) China should increase the regulation of food safety, (7) Homosexual marriage should be legalized, (8) Smoking should be banned in all public places, (9) Euthanasia should be legalized, (10) My happiest memory	Technology, love, Gaokao, failure, housewife, social progress, food safety, marriage, smoking, euthanasia, memory
17. Kassem (2021)	Are you for or against early marriage?	marriage
18. Khalil (2018)	(1) Smoking should be banned in all public places, (2) Students should not have to wear school uniforms, (3) Video games have bad effects on children, (4) Animal testing should be banned, (5) Mobile phones should be banned in schools, (6) Students should not be given homework by their teachers, (7) Money can buy your happiness, (8)Eating fast food can cause some diseases	Smoking, uniforms, video games, animal testing, phone, homework, money, happiness, fast food, diseases
19. Lustigova (2011)	- For beginner-level university classes: (1) Girls are smarter than boys, (2) Celebrities are better people than us, (3) Television makes you stupid, (4) The world will end in 2012, (5) Fashion is more important than quality, (6) The internet is just a fad, (7) Diet is better than exercise, (8) There is too much billboard advertising, (9) Little white lies are OK	Gender, celebrities, television, apocalypse, fashion, quality, internet, diet, exercise, advertisement, lies, summer, winter, phone, e-mail, technology, smoking, healthcare, president,

	sometimes, (10) Summer is better than winter, (11) The phone is more useful than e-mail, (12) Information technology should be used more in the classroom, (13) Smoking should be banned in public places, (14) There is a good system of caring for the elderly in the Czech Republic, (15) The President of the Czech Republic is doing a good job, (16) A student fee should be introduced at state-owned universities, (17) Love is more important than money, (18) It is better to be married than single - For intermediate-level university students: (1) WWII still influences life here today, (2) Friends are more important than family, (3) Without travel, we cannot appreciate life in our own country, (4) Being part of the EU helps us live better on a daily basis, (5)Without a broad CV, it is impossible to find a good job, (6) University education is vital to finding a career-based position, (7) The Czech Republic is losing its best people to the West, (8) Friends are more important than family	education fee, love, money, marriage, war, friend, family, travel, EU, job, education, job, politics, friend, family
22. Putri & Rodliyah (2021)	E.g., Dogs are better than cats, the death penalty, the ongoing educational reform	Pets, law, education, reform
24. Sabbah (2015)	Immigration, children work, fashion, old and new social traditions, private and public universities, polygamy, rumors, friendship via internet, religious sectarianism, diet, woman freedom, family cooperation, mixing at work, freedom of speech, birth control, books or visuals, endogamy	Immigration, children work, fashion, traditions, education, marriage, rumors, friendship, internet, religion, diet, woman freedom, family cooperation, work, freedom, speech, birth control, books, visuals, marriage

Appendix K Topic word frequency

No.	Word	Count	Weighted Percentage
1	education	6	4.84%
2	smoking	6	4.84%
3	technology	6	4.84%
4	marriage	5	4.03%
5	family	3	2.42%
6	freedom	3	2.42%
7	love	3	2.42%
8	money	3	2.42%
9	work	3	2.42%
10	diet	2	1.61%
11	fashion	2	1.61%
12	fee	2	1.61%
13	food	2	1.61%
14	friend	2	1.61%
15	gender	2	1.61%
16	healthcare	2	1.61%
17	internet	2	1.61%
18	job	2	1.61%
19	law	2	1.61%
20	phone	2	1.61%
21	religion	2	1.61%

22	uniforms	2	1.61%
23	women	2	1.61%
24	advertisement	1	0.81%
25	animal	1	0.81%
26	apocalypse	1	0.81%
27	arms	1	0.81%
28	attacks	1	0.81%
29	birth	1	0.81%
30	books	1	0.81%
31	celebrities	1	0.81%
32	children	1	0.81%
33	control	1	0.81%
34	cooperation	1	0.81%
35	diseases	1	0.81%
36	energy	1	0.81%
37	equality	1	0.81%
38	equity	1	0.81%
39	euthanasia	1	0.81%
40	exercise	1	0.81%
41	failure	1	0.81%
42	fast	1	0.81%
43	friendship	1	0.81%
44	games	1	0.81%
45	gaokao	1	0.81%

46	garbage	1	0.81%
47	government	1	0.81%
48	gun	1	0.81%
49	happiness	1	0.81%
50	homework	1	0.81%
51	housewife	1	0.81%
52	immigration	1	0.81%
53	leader	1	0.81%
54	lies	1	0.81%
55	mail	1	0.81%
56	memory	1	0.81%
57	party	1	0.81%
58	pets	1	0.81%
59	politics	1	0.81%
60	president	1	0.81%
61	progress	1	0.81%
62	quality	1	0.81%
63	reform	1	0.81%
64	rumors	1	0.81%
65	safety	1	0.81%
66	social	1	0.81%
67	space	1	0.81%
68	speech	1	0.81%
69	summer	1	0.81%

70	tax	1	0.81%
71	television	1	0.81%
72	terrorist	1	0.81%
73	testing	1	0.81%
74	traditions	1	0.81%
75	travel	1	0.81%
76	video	1	0.81%
77	visuals	1	0.81%
78	war	1	0.81%
79	winter	1	0.81%

Appendix L Categorisation of outcomes in language aspects

Study ID	Language aspects
1 (based on interview, perceived effects)	coherence and cohesion: coherence
	fluency
	accuracy
	vocabulary
	speech length: speaking time
2	pronunciation
	accuracy: grammatical accuracy
	fluency
	vocabulary
	interaction and involvement: interactive communication
3	general speaking proficiency: speaking ability
4	accuracy
	fluency
5 (based on ACTFL oral proficiency interview)	argumentation: opinion expression and support
	argumentation: speculation and hypotheses
	interaction and involvement: engage in extended discussions
6	general speaking proficiency: speaking ability
7	argumentation: speaking ability in giving opinion and responding to others' opinion
8 (based on Brown	argumentation: constructive speech

(2004) and Oros (2009))	argumentation: reasoning
	argumentation: rebuttal/crossfire
	fluency
	vocabulary
	grammar
	interaction and involvement: preparation
9 (based on e.g., Skehan (2009), Hyland (2005), Lee & Subtirelu (2015), Li, Chen & Sun (2015))	speech length: speech quantity (quantity of speech, and number of words produced)
	fluency
	complexity: syntactic complexity
	complexity: lexical complexity
	accuracy
	coherence and cohesion: cohesion
10 (Based on e.g., Stephen Toulmin's model of argumentation (Huh & Lee, 2014; Xargia, 2016), Qin and Karabacak (2010) and Ferretti et al. (2000))	argumentation:structural analysis
	argumentation:reasoning quality
11 (based on Hughes (2003))	pronunciation:accent
	grammar
	vocabulary
	fluency
	comprehension

12	fluency
	speech length: speaking time
13	argumentation: organization and clarity
	argumentation: use of arguments
	argumentation: use of rebuttals
	argumentation: use of examples and facts
	delivery: pronunciation, tone of voice, use of gestures, level of enthusiasm
14	general speaking proficiency: speaking skills
	vocabulary
	grammar
15 (based on Laufer (1991), Laufer and Nation's (1995) Lexical Frequency Profile, Ortega (2003))	vocabulary: lexical diversity
	vocabulary: lexical richness
	complexity: syntactic complexity
16 (based on SOLOM (Student Oral Language Observation Matrix))	comprehension
	vocabulary
	pronunciation
	grammar
	fluency
17	vocabulary
	coherence and cohesion: cohesion
	fluency

	argumentation: relevance and adequacy of content
	argumentation: adequacy and strength of arguments
18	comprehensibility and clarity: clarity
	argumentation: organisation of the speech
	vocabulary: word choices
19	general speaking proficiency: communicative competence
	speech length: speaking time
20	general speaking proficiency: communication skills
	fluency
21 (based on TOEFL iBT independent speaking rubric)	general speaking proficiency: general description
	delivery
	general speaking proficiency: language use
	comprehensibility and clarity: comprehensibility
	argumentation: topic development
22	fluency
23 (based on interview, percived effects)	fluency
	accuracy: grammatical accuracy
	vocabulary
24 (based on Brown (1997, p. 4)	pronunciation
	grammar
	vocabulary

26 (based on the impromptu speech criteria of the Toastmasters International Club)	argumentation: speech development (opening, body, close, organization, smoothness)
	argumentation: effectiveness of communication (logic, directness, enthusiasm, achievement of purpose, audience response)
	delivery: physical aspects (appearance, body language)
	delivery: voice (flexibility, volume)
	general speaking proficiency: language use (appropriate to speech, purpose and audience, good grammar, pronunciation, word selection)
27	general speaking proficiency: speaking ability
	interaction and involvement: involvement in discussion
28	fluency
	grammar
	pronunciation
	vocabulary
	comprehension
29 (based on interview, perceived effects)	general speaking proficiency: communication ability

Appendix M Narrative summary of included studies

Qualitative studies

1. Aclan and Aziz (2015) investigated the employment of the Asian Parliamentary Debate format as a pedagogical tool to enhance L2 speaking skills in the pre-debate, actual debate and post-debate phases. The data collected for analysis included semi-structured individual interviews with five experienced debate experts and a focus group interview involving students across ASEAN countries. The findings indicated an overall positive impact of using debate for developing communication skills, despite the challenges faced during the POI and rebuttal stages. Linguistically, participants reported improved speaking abilities, including coherence, fluency and accuracy of their speech, knowledge and vocabulary expansion, and longer speaking time. Affective aspects were also positively affected, as they reported that debate boosted their intrinsic and extrinsic motivation and developed self-confidence.

23. Rosas-Maldonado et al. (2020) adopted a qualitative case study to examine EFL pre-service teachers' perceptions of the potential benefits of debate pedagogy. The participants included 17 students from an English teacher education program at a university in Chile who volunteered to participate in the study. They received a 2-hour in-class debate intervention lasting five weeks and were divided into three groups of four and one group of five. Data was collected using focus group interviews after each debate activity. The results showed that participants had positive perceptions and reactions towards the debate. They reported that participating in-class debates improved their speaking fluency, and grammatical accuracy, expanded vocabulary, boosted their self-confidence and reduced speaking anxiety.

22. Putri and Rodliyah (2021) employed a qualitative case study to investigate EFL students' perception of using debate in speaking class. The study included 16 EFL students based on convenience sampling at a university in Indonesia. Participants had experienced English debate during the one-semester speaking course. After the course, ten participants filled in a questionnaire, and six were involved in the interview. The results showed that participants positively perceived integrating debate activity in the speaking class. They reported that participating in debates improves their speaking and reasoning skills, alleviates speaking anxiety, boosts confidence and prepares their speeches systematically, enhancing their fluency.

Quantitative non-randomized: the one-group pre-test-post-test pre-experimental design

17. Kassem (2021) investigated the effects of debate training on EFL learners' oral performance, ideal L2 self (the self we would like to become or motivation) and communication apprehension. The study consisted of an intact class of English majors at a university in Saudi Arabia (n = 32) lasting two months. Data collection included a speaking test and questionnaire before and after the intervention and group debating performance in every group's first, fifth, and eighth debates. Their speaking tests were rated in terms of vocabulary, cohesion, fluency, relevance and adequacy of content, and adequacy and strength of arguments with a four-point scale for each aspect. Students' level of ideal L2 self and communication apprehension was assessed via a questionnaire. The results showed a significant improvement in learners' oral performance and ideal L2 self, and their communication apprehension was reduced.

21. Noor and Tajik (2022) investigated the impacts of debate pedagogy on students' English-speaking skills at a university in Afghanistan. The study consisted of 44 sophomore students who received a six-week treatment of debate pedagogy in speaking class. Participants were evaluated through speaking tests before and after the intervention. Their speech sample was assessed based on the general description, delivery, language use, comprehensibility, and topic development with a 4-point scale for each aspect adapted from the TOEFL iBT independent speaking rubric. An online questionnaire consisting of 20 questions with a five-point scale focusing on students' perceptions of debate in speaking classes was also employed. The results showed that participants' speaking skills improved after the intervention, and they perceived using debates to develop their speaking skills and decrease speaking anxiety.

4. Baso (2016) investigated the impacts of debate pedagogy on students' EFL/ESL speaking skills in Indonesia. The participants consisted of an intact class chosen from three classes using a purposive sampling technique in a secondary school. Data were collected through a speaking test measuring accuracy and fluency. The results showed a significant difference between participants' speaking ability before and after the debate intervention.

28. Yulia and Aprilita (2017) investigated the effectiveness of implementing the British Parliamentary Debate to improve speaking ability. The participants included 12 students from an English Education Study Program at a university in Indonesia. They were evaluated via speaking tests using a five-point scale to measure fluency, grammar, pronunciation, vocabulary and comprehension. The results showed that participants improved their speaking ability after the debate intervention.

7. Desita et al. (2017) conducted classroom action research to investigate the effects of debate pedagogy in enhancing students' speaking skills. The participants were 26 students from an intact class at a secondary school in Indonesia. Data were collected using an observation checklist and field notes. The results showed that students' speaking performance improved from the first to the third cycle. They concluded that using the debate technique can enhance students' speaking ability in expressing and responding to others' opinions.

Quantitative non-randomized: the pre-test-post-test quasi-experimental design

9. El Majidi et al. (2019a) examined the effects of debate intervention on the English as an L2 speaking competence of Dutch secondary school students. The study consisted of eight intact classes from three public secondary schools in the Netherlands. Five classes (n=96) were the intervention groups, and three (n = 51) were the control groups. Most classes had three 50-minute English sessions per week, except for one intervention group that had, on average, two sessions. Both groups received regular instruction focused on four language skills. The intervention groups participated in ten weekly debates, while control groups received regular instruction to practice four language skills during that sessions further. Students were evaluated using oral opinion tasks with different topics randomly assigned to participants before and after the intervention. They had seven minutes for planning and no time limit during the performance. Their speech samples were coded as speech quantity, fluency, complexity, accuracy, and cohesion. The results of the multilevel analysis showed that the intervention groups produced more language that was fluent, accurate, coherent, and lexically sophisticated compared to the control groups.

10. El Majidi et al. (2019b) investigated the effects of debate pedagogy on various aspects of L2 argumentation. The study consisted of eight intact classes at three secondary schools in the Netherlands, with five as the intervention group (n=96) and

three as the control group (n=51). The intervention groups participated in ten weekly debates and did not receive instruction about the argumentation's structural and quality aspects. The control group received regular instruction on four language skills during the experimental group's debate session. Participants in both groups were assessed before and after the intervention by presenting oral argumentative tasks. They had 7 minutes for preparation and note-taking and had no time limits for presenting. Their speech samples were coded into argument structure and quality. The multilevel analysis results showed that the intervention groups outperformed the control groups in various aspects of oral argumentation skills regarding argument structure and quality. Participants with the debate instruction tended to reinforce their positions by presenting more backings, highlighting connections among claims through warrants, using qualifiers to temper the strength of their claims, acknowledging potential counterarguments and identifying inadequacies through rebuttals.

5. Brown (2012) investigated the effects of debate activities to enhance Russian EFL and American RFL (Russian as a foreign language) learners' oral proficiency. The study included 30 participants, 12 students from a university in Russia and 18 from the United States. They received instruction on argumentation and debate in their native language. There were eight actual debate sessions via synchronous video-conferencing, with the language used alternated each week, allowing both sides to use Russian and English, i.e., English for four debates and Russian for four debates. Both groups were evaluated through American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI) test before and after one semester's treatment. The results showed that both groups demonstrated an overall improvement in various language aspects, such as stating opinions on various topics, speculating and presenting hypotheses, and participating in the extended discourse, albeit with no significant difference between the pre-and post-test, possibly due to a small sample size.

16. Iman (2017) investigated the impacts of debate pedagogy on improving students' speaking skills in a secondary school in Indonesia. The study consisted of 48 students based on purposive sampling, and they were divided into experimental (n=24) and control groups (n=24). The experimental group received treatment using World Schools Debating format for 21 90-minute teaching sessions, while the control group had no treatment. Participants in both groups were evaluated through a

speaking test during which they were asked to choose one of the speaking topics and deliver a monologue. Their oral performance was rated based on the Student Oral Language Observation Matric (SOLOM), which consisted of different aspects of speaking skills, including comprehension, vocabulary, pronunciation, grammar, and fluency, on a five-point scale. The results showed a significant difference in speaking skills achievement between the two groups, and the experimental group performed significantly better than the control group.

15. He (2015) explored the effectiveness of a debate course compared with a traditional speaking course in enhancing university students' English speaking skills. The study included 53 students from various academic backgrounds at a Chinese university. Participants were voluntarily divided into the experimental group (n=30) and the control group (n=23). The experimental group received debate pedagogy, while the control group with the traditional teaching method used group discussions and presentations as main activities. Both courses lasted 18 weeks, with one and a half hours per week. Participants in both groups were evaluated by presenting three-minute impromptu speeches with five minutes for preparation before and after one semester's course. Their speech samples were coded into lexical richness and syntactic complexity. The results showed that students in debate class achieved more significant progress in impromptu speaking ability regarding lexical richness and syntactic complexity than in traditional speaking class.

2. Al-Tamimi (2017) investigated the effects of debate pedagogy on speaking skills and students' attitudes toward debate pedagogy at two colleges in Yemen. The study consisted of two intact groups with 70 native speakers of Arabic majoring in English. One intact class as the experimental group (n=30) were taught speaking through debate pedagogy over 16 weeks, while the control group in another intact class were taught using the regular method. Data were collected before and after the experiment using a speaking test and an attitudinal questionnaire. The speaking test included three types of speaking activities: a communication game, a reading passage, and a topic for discussion, assessing pronunciation, grammatical accuracy, fluency, vocabulary, and interactive communication. The results showed that participants who received debate instruction did significantly better in speaking tests than those who did not in the post-test. Moreover, debates positively impact students' attitudes toward speaking development.

3. Bahrani et al. (2020) investigated the effectiveness of debate pedagogy in improving L2 English speaking skills in a secondary school in Indonesia. The participants (n=50) were two intact classes chosen from three classes. One class served as the experimental group (n=26) and another as the control group (n=24). The experimental group received debate instruction over two months, while the control group received a conventional teaching method. For data collection, the researchers observed the class and filled in interview sheets before and after the treatment. The results showed that both groups had no significant difference in speaking scores in the pre-test, but the experimental group got significantly higher scores than the control group in the post-test. Moreover, the experimental group demonstrated significant improvement in speaking after debate treatment, while the control group showed no significant improvement after the conventional teaching method.

Quantitative randomised-controlled trials: the pre-test-post-test control and experimental group design

6. Dehham (2021) employed a pre-test and post-test experimental design to investigate the impact of debate pedagogy on the speaking skills of Iraqi EFL students. The participants (n = 60) were randomly selected from 5th-grade students from all schools for girls in Babylon, and they were randomly assigned into an experimental group (n = 30) and a control group (n = 30). The experimental group received debate instruction for nine weeks, while the control group followed the conventional method recommended in the textbook. The results showed no significant difference between the two groups in the pre-test, but the experimental group performed significantly better than the control group. The author concluded that debate pedagogy positively impacts Iraqi EFL students' speaking performance compared to the conventional textbook method.

11. Elmiyati (2019) adopted a pre-test-post-test control group experimental design to investigate the improvement of speaking ability through debate pedagogy and students' responses toward using debate in teaching speaking at a secondary school in Indonesia. The sample size (n = 60) was randomly chosen from six classes, and participants were randomly assigned to the experimental group (n = 30) and control group (n = 30). The experimental group was given debate treatment, while the control group had no debate treatment. Participants were assessed through a speaking

test measuring accent, grammar, vocabulary, fluency and comprehension and a questionnaire using a five-point scale. The results of the speaking tests showed that the experimental group performed better in the pre-and post-test than the control group, and students had a positive response to applying debate pedagogy in speaking class. The author concluded that the debate method is conducive to students' speaking ability.

25. Shamsudin (2017) employed a pre-test and post-test control group experimental design to investigate the impact of two teaching methods, debate and philosophy inquiry (PI), in improving ESL learners' willingness to communicate. The study consisted of 32 undergraduate students attending a speaking course at a university in Malaysia who were randomly assigned into two groups, with the experimental group receiving debate instruction and the control group receiving instruction of philosophy inquiry classroom discussion lasting eight weeks. Data were collected using the questionnaires consisting of three sections: measures of willingness to communicate, communication apprehension and communicative competence before and after the intervention. The results showed that both teaching methods improved students' willingness to communicate and self-communicative competence. However, debate pedagogy enhanced learners' willingness to communicate and self-communicative competence more effectively than the philosophy inquiry discussion group. Moreover, both groups experienced communicative apprehension, but the debate group elicited more apprehensiveness.

Mixed methods studies

8. El Majidi et al. (2018) examined the perceived effects of in-class debates on English as an L2 development among secondary school students in the Netherlands. The study spanned two years and involved seven classes (n = 134), with four classes (n = 38) taking part in the first year and three classes (n = 96) in the second year. Participants' language proficiency ranged from B1 to B2 level. They received three English sessions per week, one dedicated to debate and the remaining sessions following a regular curriculum based on course books published in the Netherlands. The debates were conducted over ten weeks. The study employed four debate formats, each with specific rules and time limits. Data collection involved a questionnaire completed by the participants, in-depth interviews with selected students, and an assessment of students' speaking performance, including constructive speech, reasoning, rebuttal/crossfire, fluency, vocabulary/variation,

grammar, and preparation. The findings showed that the participants held a positive attitude towards debates and perceived improvements in their listening, reading, writing, speaking, vocabulary, and grammar skills. They reported enhanced speaking skills, including delivering longer speeches, thinking on their feet, and improving pronunciation. The study also found a significant positive correlation between attitude towards debate and perceived effects on language development. However, there was no significant correlation between attitude and debate marks, suggesting that students of all abilities appreciated debate.

18. Khalil (2018) investigated the effects of debate pedagogy on A1 and A2 Turkish EFL students' L2 speaking anxiety, L2 speaking performance, and their perceptions toward in-class debate instruction and challenges experienced. The participants comprised 26 A1 and 24 A2 students from a Turkish secondary school. They received eight one-hour debate sessions voluntarily. Data collection included a Turkish version of the Foreign Language Speaking Anxiety questionnaire, L2 speaking tests (picture descriptions, graded based on the clarity, organisation of the speech and word choices using a 5-point scale in each aspect), researchers' filed notes and semi-structured interviews with participants. The quantitative results showed that participants' L2 speaking anxiety significantly reduced, their L2 speaking performance significantly improved, and there were no statistically significant differences between A1 and A2 participants' speaking anxiety and performance. The qualitative results reported that participants have positive perceptions toward debate instruction. At the same time, there were challenges faced by students and teachers in terms of time limitation, motivation, satisfaction with the divisions of the debate teams and the debate grades, and influences of proficiency in the selection of topics and reading materials.

26. Shan (2005) examined the impact of debate pedagogy on students' L2 English oral communication skills and the relationship with language proficiency. The participants consisted of 21 students from an intact class from an advanced oral communication course at a university in China. They were randomly assigned into groups A (n=10) and B (n=11). Group A was taught by the researcher using a debate format for the first six weeks, while Group B received instruction from a university professor using a discussion format. After six weeks, the two groups switched instruction methods to receive the alternative treatment. The data collected for analysis included three impromptu speeches, student journals, semi-structured oral

interviews, and five questionnaires. After six weeks, the results indicated that both groups showed limited progress between the pre-test and the midterm test. However, both groups demonstrated significant progress between the pre-test and the post-test after 12 weeks of intervention. Moreover, students who engaged in debate pedagogy initially and later attended the discussion format tend to experience more significant improvement in their argumentation, delivery, and general speaking proficiency than those who first engaged in discussion. Based on students' perspectives, engaging in classroom debates frequently improved their public speaking skills, increased their confidence, and reduced anxiety in speaking. Participants with higher or lower language proficiency had no significant differences between the pre-test and the post-test.

13. García-Sánchez (2020) conducted field research to investigate the effects of online in-class debate tasks for effective EFL communication. The study consisted of 126 tertiary education participants enrolled in the EFL compulsory course for 15 weeks during three consecutive academic courses from 2015 to 2017 (n = 48 in 2015, n = 37 in 2016, n = 41). Participants were divided into two teams, favour or against the debate topic, the audience and a moderator. Data collection included online schemes/scripts written by the teams, the observation of the mock debates, the final debates recorded in video and the anonymous questionnaire. The analytical observation and evaluation of scripts and recorded debates demonstrated that participants showed improvement in constructing arguments and rebuttals, as well as in pronunciation. The responses collected in the anonymous survey showed that students enjoyed the in-class debate, and their confidence and motivation to speak up in public places have been strengthened. The author concluded that students' confidence and ability in public speaking using EFL or English for Specific Purposes (ESP) could be improved through debate pedagogy.

14. García-Sánchez and Gimeno-Sanz (2022) investigated the effects of online debate through telecollaboration on developing communicative skills. The participants consisted of 25 upper-intermediate learners of ESP from two Spanish universities. Participants spent 90 minutes of online debate sessions per week lasting 15 weeks and used various digital tools to communicate and collaborate. A pre-task and post-task survey showed that telecollaborative debates are innovative and engaging ways to help increase multimodal communicative fluency in a foreign

language. Moreover, motivation and self-confidence are variables that influence the learners' performance.

24. Sabbah (2015) investigated the effectiveness of using debates in enhancing EFL students' speaking skills at a university in Palestine. The sample (n=20) was purposefully chosen from intermediate and lower-level English majors. They received debate treatment for 17 30-minute sessions lasting for a semester. They were evaluated via speaking tests. Using a five-point Likert scale, their speech samples were measured based on pronunciation, grammar and vocabulary. The results showed significant differences before and after the debate intervention. The author also provided detailed teacher guides for each debate session and concluded that debate pedagogy significantly improves students' speaking proficiency.

29. Zare and Othman (2015) examined EFL/ESL students' perceptions regarding the use of classroom debate using the British Parliamentary Debate format as a means to enhance critical thinking and oral communicative competence. Sixteen university students majoring in Teaching English as a Second Language were selected to participate. Students engaged in nine debate sessions over one semester. Data collection involved a questionnaire (21 statements using a 5-point Likert scale), reflective papers, and interviews. The authors concluded that debate pedagogy positively impacted improving students' critical thinking skills and oral communication abilities. Furthermore, the students reported additional benefits from participating in the classroom debate, including a deeper understanding of the course content, increased confidence, overcoming stage fright, and improved teamwork skills.

19. Lustigova (2011) conducted an observation study to investigate the impact of a simplified format of debates in developing EFL/ESL speaking skills. The participants included 475 intermediate and lower English proficiency students from various majors at a Czech university. Class sizes varied from 20 to 25 students. Students received a two full-semester required English course where a simplified debate format, a step-by-step approach with special attention given to the preparatory stages, was employed. The study reported that, after the intervention, students could speak in debate activities for a longer time, their anxiety was reduced, their confidence increased, and their communicative competence in English improved. The study also pointed out challenges debate teachers face regarding time constraints,

limited knowledge, students' various language proficiency and mixed academic background, and learners' occasional anxiety.

12. Fauzan (2016) conducted classroom action research where the debate technique and peer assessment were implemented to enhance students' speaking ability. The study consisted of two cycles. Each cycle involved four meetings: three meetings focused on the implementation of debate in speaking instruction, while one meeting for carrying out the speaking test. Data were collected through an observation checklist, field notes, a score sheet, and a questionnaire. The results from the pre-test and post-test showed that students' speaking ability progressed from the 'fair' category to the 'good' category. Additionally, the classroom atmosphere showed positive changes. The study also reported that students gradually developed their ability to express opinions during debate practice, improving their fluency and gaining confidence in speaking.

20. Nobuhara (2016) investigated the effects of English debate classes in leading Japanese universities' education. The study consisted of 38 students at a college in Japan who had four chances to participate in a debate activity, with 10 minutes of preparation time and 35 minutes for each debate session. Data were collected via interviews and questionnaires before and after the intervention. The results showed that students could express themselves fluently, logically and confidently in English. Students were reported to be more assertive, communicative and had more motivation to study English.

27. Tianame et al. (2019) employed classroom action research to examine the impacts of debate techniques in teaching and learning English speaking. The participants were 27 students from an intact class at a secondary school in Indonesia. They received two cycles with four meetings in total for debate training. Data collection included speaking tests in two cycles, an observation checklist and a questionnaire. The results showed that participants' speaking performance and involvement in discussion had increased from the first to the second cycle, and they responded positively to the use of debate pedagogy in teaching and learning speaking.