



Durham E-Theses

A study of the impact of the infusion method of critical thinking on Chinese students' critical thinking and academic attainment

FAN, KEJI

How to cite:

FAN, KEJI (2024) *A study of the impact of the infusion method of critical thinking on Chinese students' critical thinking and academic attainment*, Durham theses, Durham University. Available at Durham E-Theses Online: <http://etheses.dur.ac.uk/15866/>

Use policy

The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a [link](#) is made to the metadata record in Durham E-Theses
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Please consult the [full Durham E-Theses policy](#) for further details.

**A study of the impact of the infusion method of critical
thinking on Chinese students' critical thinking and
academic attainment**

Keji Fan

PhD Thesis
School of Education
Durham University
2024

Abstract

Critical thinking (CT) has been recognised as a core skill in the latest English curriculum standards for secondary schools in China. However, there is a widespread concern, particularly among Western academics, that Chinese students are not trained to develop a critical mindset. Despite this stereotypical assumption, there is little empirical evidence to decide whether this is actually the case. Misinterpreting Chinese students' difficulty in demonstrating CT as a lack of CT could result in a waste of resources. Therefore, this study first examines this common assumption through a systematic review of studies comparing the CT skills, dispositions, and styles of Chinese students with those of other nationalities.

A search of seven social science databases supplemented by other sources found 15 studies that met pre-specified inclusion criteria. Eight of these focused on students' CT skills, but their results were mixed. There is no good evidence to support the claim that Chinese students have higher or lower CT skills than students of other nationalities. Six studies on CT dispositions suggest that Chinese students were less disposed to CT, which is not the same as being weak in CT. Only one study was about CT style, indicating that Chinese students preferred information-seeking over engaging in CT. Therefore, the idea that Chinese students have weaker CT should be set aside. Additionally, almost all studies included in the review were small-scale, using weak designs. These findings suggest that the CT of Chinese students is under-studied and that more robust, larger-scale experimental studies are needed.

Most current research on CT education in China has been conducted within the higher education sector, with only a few studies suggesting that the infusion of CT shows promise in fostering CT among Chinese secondary students. However, these studies often involved small sample sizes, limiting the generalisability of their findings. Additionally, most adopted a one-group, post-test-only design. To provide more robust evidence of the efficacy of the infusion method of CT in Chinese secondary schools, a two-group randomised controlled trial with process evaluation was conducted. EnglishFusion is an intervention that infuses CT into the regular English curriculum, where CT is explicitly taught within the context of the existing curriculum. It was specifically developed for Chinese secondary school students.

Twenty-one English language teachers with 2,011 Grade 8 students from four village secondary schools in China were recruited. The randomisation occurred at the teacher level. Eleven teachers with their students (n = 1,004) were randomised to deliver the intervention once a week for three months, while the other ten teachers with their students (n = 1,007) were assigned to a business-as-usual control group. The impact of EnglishFusion was measured by differences in modified standardised CT skill tests. A process evaluation, including class observations and interviews, was also conducted to determine fidelity to the treatment and understand the mechanisms of impact evaluation findings.

The results of the trial indicate a small positive impact of infusion teaching on students' CT skills. This suggests that infusing CT into the English curriculum is a promising approach to fostering CT among Chinese secondary students. The study also demonstrates that it is feasible to train teachers to deliver EnglishFusion and that it can be incorporated into the regular curriculum without adverse effects. However, there is no evidence that improvements in CT skills translate to enhanced academic attainment outcomes. The discrepancy between improvements in CT skills and academic achievement could suggest several possibilities. One explanation is that the effects on academic attainment may require more time to manifest. Another possibility is that the cognitive load imposed on students as they simultaneously develop CT skills and learn subject content might hinder immediate academic gains. Alternatively, this discrepancy may indicate that traditional academic metrics do not fully capture or reward the cognitive growth fostered by CT.

The findings from the trial indicate that EnglishFusion appears particularly beneficial for certain groups of students. It benefits younger students more than older ones, suggesting that younger students are better positioned to gain the full benefits of CT instruction. Additionally, students from lower socioeconomic backgrounds demonstrate greater growth in both CT skills and academic achievement. Teachers' critical awareness also increased after being exposed to the training and teaching of EnglishFusion.

These findings have implications for teaching practice, educational policies, and future research. Although CT has been emphasised in the reformed curriculum, there has been no formal training for teachers on the practical aspects of delivering CT. This study demonstrates that teachers can be trained to deliver CT lessons effectively. Teacher training should emphasise student-centred pedagogies that focus on questioning skills rather than merely disseminating information.

Additionally, if developing creative and innovative thinkers is a priority in the government's ambition to become an economic and technological superpower, educational policymakers may wish to consider introducing CT at an early age, helping to build a foundation for more complex cognitive skills in later years. The school curriculum also needs to be overhauled to better balance CT and subject content. If fostering CT in schools is to be prioritised, assessment methods need to be revised to include open-ended questions that require innovative thinking, higher-order reasoning, and problem-solving skills. In China, if CT is not part of the formal examination system, it is unlikely to be taught or prioritised in schools. Given that the intervention can potentially address educational disadvantages, expanding such initiatives to underprivileged areas in China could also help bridge educational gaps, ensuring that all students, regardless of socioeconomic background, have the opportunity to develop these essential skills.

Future research could explore the long-term impact of infusing CT and investigate how best to integrate CT into different stages of the educational process.

Table of contents

Abstract	i
List of tables	viii
List of figures	x
List of abbreviations	xi
Declaration	xii
Statement of copyright	xiii
Acknowledgements	xiv
Section I Introduction	1
Chapter 1 The rationale of the research	4
1.1 Background	4
1.2 Aims of the study.....	7
1.3 Research questions	8
1.4 Significance of the study	8
Chapter 2 Critical thinking — the debates.....	10
2.1 What is critical thinking?.....	10
2.2 Can critical thinking be taught?.....	19
2.3 How can critical thinking be taught?.....	21
2.4 How is critical thinking measured?	27
2.5 Chapter summary.....	32
Chapter 3 The critical thinking of Chinese students	33
3.1 Common perceptions of Chinese students’ critical thinking.....	33
3.2 Possible explanations for the stereotype.....	35
3.3 Challenges to the common perceptions	38
3.4 Strategies for improving critical thinking of Chinese students	40
3.5 Chapter summary.....	45
Chapter 4 Critical thinking education in China	46
4.1 The development of policy on critical thinking education	46
4.2 Critical thinking education in China.....	48
4.3 Teacher training for critical thinking education	49
4.4 Challenges in implementing critical thinking in schools in China	51
4.5 Chapter summary.....	52
Section II Research design and methods	53
Chapter 5 Systematic review	54

5.1 Research aim and questions.....	54
5.2 Rationale for a systematic review.....	54
5.3 Searching strategy	56
5.4 Screening.....	57
5.5 Data extraction.....	59
5.6 Quality assessment	60
5.7 Synthesis.....	63
5.8 Chapter summary.....	64
Chapter 6 Primary research.....	65
6.1 Research aims and questions	65
6.2 The pilot study.....	66
6.3 The intervention.....	66
6.4 Trial design.....	84
6.5 The sample	86
6.6 Outcome measures.....	89
6.7 Analyses	104
6.8 Process evaluation	108
6.9 Ethics	112
6.10 Chapter summary.....	113
Section III Results of the systematic review	114
Chapter 7 Results of the structured review	115
7.1 Results of the search.....	115
7.2 Strength of evidence of included studies	118
7.3 How do Chinese students' critical thinking skills compare with those of other nationalities?.....	118
7.4 How do Chinese students' critical thinking dispositions compare with those of other nationalities?.....	124
7.5 How do Chinese students' critical thinking styles compare with those of other nationalities?.....	128
7.6 Chapter summary.....	130
Section IV Results of the primary research	131
Chapter 8 Results of the pilot study.....	132
8.1 Revisions for the main study	132
8.2 Impact evaluation	134
8.3 Chapter summary.....	136
Chapter 9 Impact of EnglishFusion on critical thinking.....	138

9.1 The sample	138
9.2 Characteristics of students	142
9.3 Characteristics of teachers	145
9.4 Does EnglishFusion improve Chinese secondary students' critical thinking skills?	146
9.5 Does EnglishFusion have a differential impact on the critical thinking skills of sub-groups of students?	152
9.6 Does training and teaching EnglishFusion alter teachers' critical awareness and attitudes towards teaching critical thinking?	158
9.7 Chapter summary.....	160
Chapter 10 Impact of EnglishFusion on academic attainment	162
10.1 Does EnglishFusion improve Chinese secondary students' academic performance?	162
10.2 Does EnglishFusion have a differential impact on the academic attainment of sub-groups of students?	168
10.3 Chapter summary.....	172
Chapter 11 Results of process evaluation	174
11.1 Fidelity to implementation	174
11.2 Teacher training and preparation.....	175
11.3 Students' opinions on EnglishFusion	176
11.4 Teachers' views on EnglishFusion.....	181
11.5 Perceptions of the impact of EnglishFusion on critical thinking skills	182
11.6 Perceptions of the impact of EnglishFusion on academic attainment	185
11.7 Challenges to successful implementation.....	187
11.8 Conditions for successful implementation	189
11.9 Summary of this chapter.....	191
Section V Discussion and conclusions	192
Chapter 12 Discussion	193
12.1 Summary of findings	193
12.2 Limitations.....	199
Chapter 13 Implications and conclusions	202
13.1 Implications of the systematic review	202
13.2 Implications and recommendations of the trial	203
13.3 Conclusions	207
References	209
Appendices.....	236
Appendix A. Search syntax and results in databases.....	236

Appendix B. Data extraction tables.....	237
Appendix C. An example of the teaching materials.....	247
Appendix D. The pre-and post-CT tests.....	256
Appendix E. Teacher questionnaire.....	282
Appendix F. An example of class observation notes.....	286
Appendix G. Ethical approval	308
Appendix H. Sub-group analyses results.....	310

List of tables

Table 2.1 Typology of Ennis’s (1989) critical thinking teaching methods.....	22
Table 5.1 A “sieve” to assist with quality assessment.....	63
Table 6.1 Summary of EnglishFusion content at two stages.....	68
Table 6.2 Practice with conclusions and assumptions.....	73
Table 6.3 Results of teacher randomisation	87
Table 7.1 Summary of strength of evidence for all studies (N = 15)	118
Table 7.2 Summary of results on literature comparing critical thinking skills (n = 8).....	119
Table 7.3 Summary of results on literature comparing critical thinking dispositions (n = 6).....	125
Table 7.4 Summary of results of literature comparing critical thinking styles (n = 1).....	129
Table 8.1 Summary of main differences between the pilot study and the main trial	132
Table 8.2 Comparison of pre-test critical thinking scores between experimental and control groups (N = 122).....	134
Table 8.3 Comparison of post-test critical thinking scores between experimental and control groups (N = 122).....	135
Table 8.4 Comparison of gain in critical thinking scores between experimental and control classes (N = 122).....	135
Table 9.1 Characteristics of the four participating schools	139
Table 9.2 Percentage of birth sex in experimental and control groups (N = 2,055).....	142
Table 9.3 Percentage of ethnicity in experimental and control groups (N = 2,055).....	142
Table 9.4 Comparison of household possessions in experimental and control groups (N = 2,055)	143
Table 9.5 Parental involvement in experimental and control groups (N = 2,055)	144
Table 9.6 Comparison of teachers’ age between experimental and control groups (N = 21).....	145
Table 9.7 Comparison of English teaching experience between experimental and control groups (N = 21).....	145
Table 9.8 Comparison of pre-test critical thinking scores between experimental and control groups (N = 2,011).....	146
Table 9.9 Comparison of post-test critical thinking scores between experimental and control groups (N = 2,011).....	147
Table 9.10 Comparison of gain in critical thinking skills scores between experimental and control groups (N = 2,011)	147
Table 9.11 Comparison of pre-test critical thinking scores for missing cases between experimental and control groups (N = 44).....	149
Table 9.12 Comparison of pre-test critical thinking scores between missing and completed cases .	150
Table 9.13 Regression results predicting students’ post-test critical thinking scores.....	151
Table 9.14 Coefficients for the model predicting post-test critical thinking scores	152

Table 9.15 Comparison of impact on gain in critical thinking by demographic characteristics* .	153
Table 9.16 Comparison of impact on gain in critical thinking by household possessions	154
Table 9.17 Comparison of impact on gain in critical thinking by cultural capital	155
Table 9.18 Comparison of impact on gain in critical thinking by parental involvement	156
Table 9.19 Comparison of impact on gain in critical thinking by schools	156
Table 9.20 Comparison of impact on gain in critical thinking by prior academic attainment	157
Table 9.21 Comparison of impact on gain in critical thinking by prior critical thinking scores ...	158
Table 9.22 Teachers' agreement on the trustworthiness of research findings (N = 21).....	158
Table 9.23 Teachers' attitudes towards teaching critical thinking (N = 21)	160
Table 10.1 Comparison of pre-test academic scores between experimental and control groups ..	163
Table 10.2 Comparison of post-test academic scores between experimental and control groups.	163
Table 10.3 Comparison of gains in academic scores between experimental and control groups..	164
Table 10.4 Results of sensitivity analysis of academic attainment	165
Table 10.5 Comparison of pre-test academic scores of students missing the post-tests.....	166
Table 10.6 Regression results predicting students' post-test academic scores	167
Table 10.7 Coefficients for the model predicting post-test academic scores	168
Table 10.8 Comparison of impact on gain in academic scores by demographic characteristics* .	169
Table 10.9 Comparison of impact on gain in academic scores by home possessions.....	169
Table 10.10 Comparison of impact on gain in academic scores by cultural capital.....	170
Table 10.11 Comparison of impact on gain in academic scores by parental involvement.....	171
Table 10.12 Comparison of impact on gain in academic scores by schools	171
Table 10.13 Comparison of impact on gain in academic scores by other sub-groups.....	172

List of figures

Figure 6.1 An extract of the student handout for Lesson 5	77
Figure 6.2 An example of lesson slides for an activity in Lesson 3	78
Figure 6.3 An extract of the lesson plan for Lesson 6.....	80
Figure 6.4 Sample item for argument evaluation.....	91
Figure 6.5 Sample item for assumption identification	92
Figure 6.6 Sample item for deduction skill.....	92
Figure 6.7 Sample item for inferential skill	93
Figure 6.8 Sample item for interpreting information	93
Figure 7.1 PRISMA flow diagram	117
Figure 9.1 The CONSORT flow diagram	140

List of abbreviations

APA	American Psychological Association
ASSIA	Applied Social Sciences Index & Abstracts
CCTDI	California Critical Thinking Disposition Inventory
CCTST	California Critical Thinking Skills Test
CCTT	Cornell Critical Thinking Test
CONSORT	Consolidated Standards of Reporting Trials
CT	Critical Thinking
ERIC	Education Resources Information Center
ES	Effect Size
HCTA	Halpern Critical Thinking Assessment
IQ	Intelligent Quotient
MoE	Ministry of Education
MSLQ	Motivated Strategies for Learning Questionnaire
NNTD	Number of counterfactual cases Needed to Disturb the finding
OECD	Organisation for Economic Co-operation and Development
Ofsted	Office for Standards in Education, Children's Services and Skills
PBL	Problem-Based Learning
PISA	Programme for International Student Assessment
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RCT	Randomised Controlled Trial
SD	Standard Deviation
SES	Socioeconomic Status
SR	Systematic Review
STEM	Science, Technology, Engineering, and Maths
UFCTI	University of Florida Critical Thinking Inventory
WGCTA	Watson-Glaser Critical Thinking Appraisal

Declaration

I declare that this thesis is my own work and has not previously been submitted elsewhere for any other qualification or degree.

Part of this thesis, the systematic review and the randomised controlled trial, has been published in peer-reviewed journals and a scholarly book. Below are the references of the publications:

Fan, K. (2024). Can the Infusion Teaching of Critical Thinking Improve Chinese Secondary Students' Critical Thinking and Academic Attainment? Findings from a Randomised Controlled Trial. *Thinking Skills and Creativity*, 53. Article 101597. <https://doi.org/10.1016/j.tsc.2024.101597>

Fan, K. (2024). Are Chinese students deficient in critical thinking? In S. Gorard & N. Siddiqui (Eds.), *An International Approach to Developing Early Career Researchers: A Pipeline to Robust Education Research*. London: Routledge.

Fan, K., & See, B. H. (2022). How do Chinese students' critical thinking compare with other students?: a structured review of the existing evidence. *Thinking Skills and Creativity*, 46. Article 101145. <https://doi.org/10.1016/j.tsc.2022.101145>

Statement of copyright

The copyright of this thesis rests with the author. No quotation from it should be published without the author's prior written consent and information derived from it should be acknowledged.

Acknowledgements

I would like to express my gratitude to my PhD supervisors, Professor Beng Huat See and Professor Stephen Gorard, for their ongoing support and invaluable guidance throughout this project. I am also sincerely thankful to Professor Nadia Siddiqui for her insightful assistance. Their support has made my PhD journey both smooth and inspiring.

I extend my heartfelt thanks to the participants who took part in this study. Without their involvement, the successful completion of this ambitious large-scale trial would not have been possible.

Lastly, I would like to thank my parents, friends, and partner for their constant encouragement and support.

Section I Introduction

This thesis examines the teaching and learning of critical thinking (CT) in countries where English is not the first language, focusing particularly on China. With the rapid proliferation of information and its easy accessibility through social media platforms such as Google, Facebook, Twitter (X), blogs, instant messaging, and TikTok, students are continually bombarded with various types of information, including fake information, disinformation, misinformation, and propaganda. In the current climate, what students need is not more information, but the ability to sift through it, interpret reported data, evaluate evidence, critically appraise the quality of such evidence, and discern what is believable and what is not (Horn & Veermans, 2019). Such skills are particularly essential for young people in the 21st century.

Over the last three decades, policymakers in China have reportedly increasingly recognised the importance of educating citizens to think critically, creatively, and innovatively to compete in the global economic and technological arena. Educational reforms in higher education aimed at fostering CT have led to a rise in the number of courses and programmes on CT in China (e.g. Dong, 2015; Jiang, 2013; Xu, 2019). However, the introduction of CT teaching in China has encountered many challenges. First, it is primarily implemented in the higher education sector, with students in primary and secondary schools rarely exposed to CT teaching (Dong, 2015). This was contrary to the suggestion that CT should be introduced early in education (Kuhn, 1999). Secondly, teachers in China have not been adequately prepared to teach CT. They continue to face pressure to prepare students for high-stakes assessments (Tan, 2020). Furthermore, most teacher training has focused solely on the theory and importance of CT, lacking sufficient high-quality training with hands-on activities to enable them to effectively teach CT in the classroom (Yan, 2012).

Perhaps due to these challenges, Chinese students often lack familiarity with CT and apparently struggle to understand its essence and demonstrate it in their work. Chinese students studying overseas have frequently been depicted as overly passive recipients of information, unquestioning and uncritical. This thesis challenges that view by first evaluating the evidence for the common assumption regarding the lack of criticality among Chinese students. This is achieved through a systematic review, allowing for an

in-depth analysis of existing literature (Siddaway, Wood, & Hedges, 2019) and an evaluation of the trustworthiness of findings (Gorard, 2013). Subsequently, it conducts a randomised controlled trial to test the effectiveness of the infusion method of teaching CT in the English curriculum (i.e. EnglishFusion) in Chinese classrooms.

This thesis comprises five sections. Section I serves as the introductory section and consists of four chapters. Chapter 1 introduces the background and rationale for the study, research purposes, and research questions, while also presenting the research significance. Chapter 2 discusses ongoing debates in the field of CT, including various definitions of CT, whether CT can be taught, and how it should be taught and measured. Chapter 3 focuses on CT in the Chinese context, describing the typical portrayal of Chinese students' criticality (or lack thereof) and offering possible explanations for this common perception. The discussion in Chapter 3 justifies the need to establish evidence for claims about Chinese students' CT. Chapter 4 explores the development of CT education in China over the last two and a half decades, alongside pertinent policies. It shows the present landscape of CT instruction within schools and the preparation of teachers to deliver CT in schools, setting the context for the thesis and justifying the use of the infusion method of CT teaching in the English curriculum in Chinese secondary schools.

Section II presents the research design and methods used in the thesis and comprises two chapters. Chapter 5 justifies conducting a systematic review (SR) and describes the SR process from searching, screening, extracting, and evaluating to synthesising. The SR is conducted first because, without evidence that Chinese students lack CT skills, resources allocated to designing interventions to improve these skills would be wasted. Chapter 6 details the methods and design for the primary research, including the intervention, sample, outcome measures, analyses, process evaluation, and ethical considerations. This chapter also reports details of a pilot study that was conducted prior to the main trial. Changes in design, intervention and test instruments for the main trial were also clarified.

Section III presents the results of SR. There is only one chapter. Chapter 7 addressed the research question about Chinese students' CT in comparison to those from other

nationalities. Three dimensions of CT were compared and analysed, including skills, dispositions and styles.

Section IV reports the results of the primary research and comprises four chapters. Chapter 8 discusses the pilot study results and how feedback on the intervention and test instruments informed the main study. The subsequent three chapters report the results of the main trial. Chapter 9 examines the impact of EnglishFusion on CT skills, while Chapter 10 focuses on academic attainment. Chapter 11 presents the process evaluation, which assesses implementation fidelity, collects the perspectives of teachers and students and identifies the challenges and conditions for successful implementation.

Section V is the final section of the thesis and consists of two chapters. Chapter 12 summarises the main research findings, synthesising results from both SR and the cluster-randomised controlled trial. It discusses limitations and suggests avenues for future studies. Chapter 13 considers the implications of these findings and presents the conclusions of the whole project.

Chapter 1 The rationale of the research

1.1 Background

Critical thinking (CT) has been identified as a core skill for the increasingly complex and globalised economies of the 21st century (Organisation for Economic Co-operation and Development [OECD], 2018a). Rapid technological and communication advances mean that people can now have immediate and easy access to information. CT enables individuals to sift through vast amounts of information, distinguishing between reliable and valid information and misinformation (Horn & Veermans, 2019). It supports the development of problem-solving skills and helps individuals make informed decisions. Additionally, it helps in developing communication skills, enabling individuals to present their ideas and arguments coherently. These skills are particularly important in the complex modern world. They offer a layer of protection for young people against fake news, allowing them to judge for themselves what is true, what is propaganda, and what is not. Associated with CT is the ability to ask critical questions and seek evidence to support claims or beliefs. While CT is not a new concept, it is deemed more essential in contemporary societies.

Attempts to develop CT in some societies may face challenges due to cultural norms, political systems, or educational structures, but there are opportunities for its promotion and development. China is an example where economic reform and the open-door policy in the late 1970s led to the growth of the high-tech industry and an influx of foreign investments, necessitating the development of skills and talents to stay competitive in the global economy. Consequently, China's education system needs to be reformed as well. Traditionally, the Chinese education system emphasises rote memorisation and regular testing. However, there is now a growing recognition of the importance of fostering CT skills to prepare its citizens for the modern world. Coupled with the proliferation of information and the widespread use of fake news, this forms a perfect background for research into the teaching of CT and the preparation of teachers in delivering the new curriculum.

China's education system has undergone radical reform in the last two decades. One of the key aims of this reform was to shift the focus of teaching and learning away from the mere acquisition of basic knowledge and skills towards the development of CT,

creativity, and problem-solving abilities (Cheng, 2010). As part of China's plan to become a global economic and technological superpower, its leaders pledged to modernise the education system with the publication of the report titled Outline of China's National Plan for Medium and Long-term Education Reform and Development (2010-2020) (People's Republic of China, 2010). The report focused on cultivating well-rounded citizens and improving the education system. While it recognises the importance of learning with thinking, it provides insufficient information on the desired thinking skills and how to enhance them.

In April 2022, the Ministry of Education (MoE) of the People's Republic of China issued a new curriculum plan and standards for primary and secondary schools, emphasising CT (MoE, 2022a, 2022b). These have been implemented since September 2022. The new English curriculum standards for secondary schools (MoE, 2022b) emphasises skills such as interpretation, analysis, evaluation, inference, explanation, deduction, assumption, and self-regulation, all of which are elements of CT.

Despite policy efforts to foster CT in schools, Chinese teachers were not adequately prepared to teach CT in the classroom. They still faced the pressure of preparing students for high-stakes assessments (Tan, 2020). Additionally, there was a lack of sufficient high-quality teacher training with hands-on activities to facilitate teaching CT in class (Yan, 2012). Most teacher training focused on the theory and importance of CT, posing a barrier to effective teaching of CT in schools. Hence, there is a dearth of studies focusing on CT among schoolchildren (Fung, 2017).

Some schools have explored ways of embedding CT in the regular curriculum (Li, 2017; Zhai, 2015). One of these is the infusion of CT into the regular curriculum, where subject content is used, and the improvement of CT is made explicit to students (Ennis, 1989). This infusion method is deemed appropriate in the Chinese context and is more likely to be accepted by schools as it integrates CT into the existing curriculum using the textbooks and materials already in use, rather than introducing a standalone course (Tan, 2020). It also fits well with school timetabling. However, many previous studies on the impact of the infusion method have methodological limitations. Small sample sizes limit the generalisability of their findings. Many studies recruited no more than

100 students (e.g. Bağ & Gürsoy, 2021; Dong, 2017; Lin, 2014; Marin & Halpern, 2011; Zohar & Tamir, 1993). In addition, many studies do not carefully consider the fidelity to treatment and diffusion problems (Toomey et al., 2020). For instance, in the quasi-experimental study conducted by Bağ and Gürsoy (2021), it is unknown whether the instructor taught both experimental and control classes. There might be a diffusion problem where the teacher unconsciously taught CT-relevant content to the control class. On the other hand, if there were two teachers, any impact might be due to teacher differences and cannot be exclusively attributed to the intervention.

Similarly, in Hu et al.'s (2011) study, experimental and control students were in the same class, and in Lin's (2014) study, the same teacher taught both the infusion and traditional English courses. This considerably increased the problem of diffusion, but neither study made any efforts to prevent it. To determine whether the infusion of CT in the regular curriculum has any impact on improving Chinese secondary students' CT skills and academic attainment, my thesis employs a randomised controlled trial design.

The education reforms of the last two decades have also seen an increasing number of Chinese students studying overseas in Western English-speaking universities (OECD, 2013). There is a growing interest in the learning skills and dispositions of Chinese students. One common Western stereotype is that Chinese students are not critically aware (Song, 2014; Xu, 2021). For instance, Guo and O'Sullivan (2012) observed that Chinese students were not familiar with CT and misunderstood it as negative thinking. Chinese students are often perceived as passive recipients of knowledge, lacking skills in analysis and evaluation (Lucas, 2019). Their learning has been described as superficial (Watkins & Biggs, 1996), focusing on memorisation rather than interpretation or analysis.

In many cases, however, the judgement of Chinese students' CT ability is based on subjective impressions, which are notoriously unreliable. A more reliable evaluation of CT skills would be the use of standardised tests (Gorard, See, & Siddiqui, 2017). Few studies have used standardised tests to measure Chinese students' CT. Furthermore, the few that did only measured the CT skill levels of students without any comparison. Without comparisons with students from other nationalities, it is not possible to

conclude whether Chinese students have higher, lower, or comparable CT skills to similar students in Western democracies (Gorard, 2013). The assumption that Chinese students lack CT implies a comparison. What are we comparing Chinese students' CT with, whose CT are we comparing, and what does the norm look like? Most research into students' CT lacks a comparator, yet makes bold claims about the low levels of CT skills of Chinese students. This is absurd, yet widely accepted. For this reason, this thesis will review studies that compare Chinese students' CT with that of other nationalities using validated standardised tests. The systematic review will first establish the evidence on Chinese students' CT and determine whether it is indeed poorer than that of students in other countries.

1.2 Aims of the study

One of the objectives of this thesis is to establish the evidence of Chinese students' CT levels through a systematic review. This review gathers and critically appraises international evidence to reach a well-informed conclusion about the CT levels of Chinese students. The advantage of a systematic review is that it consolidates all relevant evidence in one place, allowing for a comprehensive assessment of each study's quality and facilitating a more objective judgement. For this review, I will include only studies that feature a comparison group. Since comparisons are often made between Chinese students and students of other nationalities, particularly from English-speaking countries, it is vital that the studies included in this review have appropriate comparators.

A key objective of this thesis is to evaluate the impact of the infusion method of delivering CT (i.e. EnglishFusion) on Chinese students' CT skills and academic attainment. Additionally, this thesis contributes to the debate on whether CT is a product of Western philosophy and thus cannot be taught in non-Western contexts. This will involve examining whether CT can indeed be taught effectively in non-Western contexts, countering the prevailing notion that the Chinese education system emphasises rote memorisation and that teachers' primary role is to disseminate information.

1.3 Research questions

Aligned with the above research purposes, the following research questions are raised:

1. What is the evidence on Chinese students' critical thinking compared with students of other nationalities?
2. Can critical thinking skills be taught to Chinese secondary students who are not traditionally exposed to critical thinking?
 - 3a. Does EnglishFusion improve Chinese secondary students' critical thinking skills?
 - 3b. Does EnglishFusion have a differential impact on the critical thinking skills of sub-groups of students (by age, birth sex, ethnicity, prior academic attainment, prior critical thinking skills, schools, parental involvement in children's education, and home background)?
 - 4a. Does EnglishFusion improve Chinese secondary students' academic performance?
 - 4b. Does EnglishFusion have a differential impact on the academic attainment of sub-groups of students (by age, birth sex, ethnicity, prior academic attainment, prior critical thinking skills, schools, parental involvement in children's education, and home background)?
5. Does training and teaching EnglishFusion alter teachers' critical awareness and attitudes towards teaching critical thinking?

To address research question (RQ) 1, a systematic review was conducted. RQs 2 to 5 were addressed using a cluster randomised controlled trial.

1.4 Significance of the study

This thesis contributes to current debates about the CT of Chinese students. While previous studies have explored this area, the results have been inconclusive. The research in this thesis consists of two main strands, each with its own significance.

This systematic review differs from previous ones in two key ways. Firstly, while previous reviews have examined this issue by focusing on Asian students in general (e.g. Indra, 2019; Salsali, Tajvidi, & Ghiyasvandian, 2013), this review specifically focuses on Chinese students. Secondly, unlike previous reviews (e.g. Huang, 2019; Tian & Low, 2011) which provided mere summaries of research findings without

considering the quality of evidence, this new review will only consider robust studies with comparators, using experimental or quasi-experimental designs. Additionally, each study will be assessed for the quality of evidence based on study design, scale, and threats to validity, thus providing more confidence in the findings. This will represent the best evidence available regarding the CT abilities of Chinese students.

The second strand is the primary research, which involves a randomised controlled trial (RCT) of the infusion method of teaching CT in secondary schools in China. Most current research on CT education in China is conducted in the higher education sector (e.g. Cheng, Huang, Yang, & Chang, 2020; Cui & Teo, 2023; Dong, 2017; Yuan, Kunaviktikul, Klunklin, & Williams, 2008; Zhang et al., 2017), and only a few focused on secondary school students in China (e.g. Fung, 2017; Ku, Ho, Hau, & Lai, 2014; Wang, Chen, Lin, Huang, & Hong, 2017). To the best of my knowledge, this is the first study to employ an RCT to test the infusion of CT on a large sample of secondary school students in China. The findings of this study will contribute to ongoing efforts to enhance CT education in Chinese secondary schools. The study also has the potential to address the challenges faced by Chinese educators in fostering CT skills amidst high-stakes assessments and limited teacher training. As part of the research, a suite of teaching and learning resources will be developed for use in schools, and teachers will receive training on how to embed CT teaching into the existing curriculum. Therefore, this research offers professional development opportunities for teachers.

If found to be effective, the infusion method of CT would offer a practical way of introducing CT teaching into schools with minimal disruption to the existing curriculum. The research holds significance in developing students' CT competency and addressing the Chinese government's ambition for its citizens to remain competitive in the global education landscape.

Chapter 2 Critical thinking — the debates

Critical thinking (CT) is a challenging concept to define (Byrne, 1994; Fisher, 2011; Nilson, 2021). There is no universally agreed-upon definition as it is a complex and multifaceted concept that involves a variety of cognitive skills and dispositions. Different disciplines and educational philosophies emphasise different aspects of CT, leading to various interpretations of the concept. In this chapter, I will discuss the different definitions and models of CT and address the question as to whether CT can be taught. Some argue that CT is like intelligence – you either possess it or you do not. Others view CT as a set of skills that can be developed. If this is the case, how can CT be taught? This chapter will explore these debates. A crucial aspect of these discussions is the question of measurement — how do we measure CT, and can it even be measured?

2.1 What is critical thinking?

CT is a concept with multiple interpretations (Johnson, 1992; Moore, 2013). It draws from fields such as philosophy, cognitive psychology, and education (Lai, 2011). Each discipline offers its own perspective on what constitutes CT. The philosophical roots of defining CT can be traced back to early Greek philosophy (Facione, Sánchez, Facione, & Gainen, 1995). Socrates, for example, emphasised the necessity of critical inquiry for a meaningful existence with his famous assertion, “The unexamined life is not worth living.” His method of questioning, known as “Socratic questioning,” has laid the foundation for CT teaching. Socrates demonstrated that even those in positions of power and influence could possess deeply flawed reasoning, and thus, people should not be too ready to accept ideas. They should remain sceptical, seek relevant evidence and examine common beliefs and assumptions logically (Paul, Elder, & Bartell, 1997).

Following Socrates, Plato and Aristotle further developed the tradition of CT. Plato, in his dialogues, meticulously documented Socrates’ inquiries and philosophical explorations, perpetuating the legacy of critical inquiry. Aristotle, on the other hand, emphasised the need to discern between the superficial claims and the deep nature of things (Paul et al., 1997). He advocated for comprehensive and systematic thinking to enable individuals to look beyond the surface. These Greek philosophers contributed to this intellectual tradition by challenging prevailing beliefs and highlighting the importance of scepticism in uncovering truth. Collectively, they set the agenda for CT.

John Dewey, a pragmatic philosopher, was also among the earliest individuals to coin the term “critical thinking” where he referred to CT as “reflective thinking”. According to him, reflective thinking is an “active, persistent, and careful consideration of a belief or supposed form of knowledge in the light of the grounds which support it and the further conclusions to which it tends” (Dewey, 1909, p. 9). By defining CT as an **active** process, Dewey stressed that people should be responsible for their thinking, such as raising questions and searching for relevant information themselves. This contrasts with passive thinking, where people merely receive information from others without questioning. By considering CT as a **persistent** and **careful** process, Dewey also recognised that critical thinkers do not simply jump from received information to a conclusion without warranting careful thought. CT is consistently required, and each step of thinking should be made carefully and logically. Reflective thinking, as Dewey defined it, requires “the voluntary effort to establish belief upon a firm basis of evidence and rationality” (Dewey, 1933, p. 9).

Many philosophers, influenced by John Dewey, incorporated the idea of reflective thinking into the conceptualisation of CT. McPeck (1981), for instance, proposed that CT involves the tendency and skill to engage in activities with reflective scepticism. Likewise, Ennis (1987) defined CT as a form of reasonable, reflective thinking aimed at determining what one should believe or do. Lipman (2003) argued that the outcomes of thinking do not have to be externalised, and that the process of thinking should be focused. He identified three main components of CT: using suitable criteria, self-reflection, and a sense of context (Lipman, 1988). Facione (1990) presented a more specific definition of CT as “purposeful, self-regulatory judgment which results in interpretation, analysis, evaluation, and inference, as well as explanation of the evidential, conceptual, methodological, criteriological, and contextual considerations upon which that judgment is based” (p. 3). However, despite the variations in defining CT in the philosophy field, there is general agreement that it involves logical reasoning, evaluation of arguments, and rigorous analysis (Dwyer, Hogan, & Stewart, 2014).

Cognitive psychologists, on the other hand, argue that philosophical definitions of CT are too idealistic, presuming how people would behave under the “best” conditions

(Sternberg, 1986). They have sought to frame CT from a different perspective, considering its nature in authentic contexts (Black, 2007). According to psychologists, CT is a construct that cannot be directly observed (Bailin, 2002). To define CT in a way that can be observed and measured, cognitive psychologists tend to focus on the outcomes of CT, including a set of behaviours and skills of critical thinkers (Halpern, 1999; Lai, 2011). Sternberg (1986), for example, contended that CT includes mental processes, techniques, and representations that individuals employ for problem-solving, decision-making, and acquiring new knowledge. Similarly, Halpern (1998) stated that CT is about the application of cognitive skills for desirable outcomes. Willingham (2008) listed a collection of critical thinkers' behaviours, including "seeing both sides of an issue, being open to new evidence that disconfirms your ideas, reasoning dispassionately, demanding that claims be backed by evidence, deducing and inferring conclusions from available facts, solving problems, and so forth" (p. 21).

Educational psychologists focus on learning and instruction, including how CT can be developed (Barnett & Davies, 2015). Bloom's taxonomy is sometimes recognised as CT, involving skills of memorising, understanding, applying, analysing, evaluating, and creating (Krathwohl, 2002). From an educational perspective, CT is seen as a desirable goal that can be taught, measured, and applied (Dhakal, Watson Todd, & Jaturapitakkul, 2023).

Although there are various conceptions of CT, there is no consensus on what exactly CT is. What we do know, however, is that CT is a complex concept with different interpretations across disciplines (Black, 2007). CT skills also overlap with other skills such as creativity and metacognition, which makes it even more challenging to define. For instance, Fisher (2011) coined the term "critico-creative thinking" that links both creativity and CT to highlight the positive and imaginative aspects of CT. The term "critical" may sometimes be misunderstood as mere criticism, which can seem quite negative. However, CT involves an imaginative dimension as it allows individuals to consider different stances and imagine alternative scenarios (Fisher, 2011). This term is in line with Paul and Elder's (2006) argument that CT without creativity would amount to mere scepticism and negativity.

CT also has connections to meta-cognition, which is often referred to as thinking about thinking (Kuhn, 1999; Schraw, Crippen, & Hartley, 2006). Kuhn (1999) perceives CT as a facet of meta-cognition, comprising three components: meta-cognitive knowing, meta-strategic knowing, and epistemological knowing. Meta-cognitive knowing is the basic level, concerned with declarative knowledge, while meta-strategic knowing is more advanced, focusing on the monitoring and evaluation of one's thinking. Epistemological knowing addresses broader questions about knowledge, such as how knowledge is produced. On the contrary, Van Gelder (2005) and Willingham (2008) view CT as an overarching term, with meta-cognition as merely a subset. All of these indicate that it is a difficult job to define CT.

Since a single definition can hardly encompass the complexity of CT, the concept should not be perceived in isolation (Bennett, Faltn, & Wright, 2003). As Black (2007) contends, CT can be interpreted from two dimensions: cognitive skills and affective dispositions. Both dimensions serve to enhance the awareness of reality (Sievers, 2001). Additionally, another facet of CT is the thinking style, which focuses on how CT is exercised when people solve problems (Lamm, 2015). The following sections present different models of CT skills, dispositions, and styles.

Critical thinking skills

There are various models of CT skills. Each offers unique perspectives and frameworks to understand and cultivate CT. In this section, I will present three prominent models of CT skills, highlighting their key principles and contributions to the understanding of essential cognitive skills.

The Ennis model of CT abilities

One of the leading definitions proposed by Ennis (1987) defines CT as “reasonable reflective thinking that is focused on deciding what to believe or do” (p. 10). Ennis (2015) suggested fifteen abilities or skills associated with an ideal critical thinker who is capable of drawing well-founded conclusions. These can be summarised as follows:

- Analysing arguments: critically examining the reasoning and evidence presented in arguments to evaluate their validity and soundness.

- Evaluating sources: judging the credibility of sources to discern reliable information from unreliable or biased sources.
- Making inferences and judgements: making and judging both inductive and deductive inferences and arguments, as well as forming and evaluating judgements based on available evidence.
- Clarifying concepts: asking and answering clarification questions, defining terms, and identifying and handling equivocation appropriately.

Paul and Elder's model of CT skills

According to Paul and Elder (2010), CT entails independent, disciplined, self-monitored, and self-adjusting thought processes, and it requires trustworthy standards to analyse and evaluate the structures inherent in thinking. In this framework, people's thinking is assumed to be naturally biased and distorted (Elder & Paul, 2020), and thus CT must be systematically cultivated. This requires rigorous intellectual standards. These nine intellectual standards are:

- Clarity: ensuring one's argument is clear and free from ambiguity.
- Accuracy: points made must be accurate and consistent.
- Precision: providing detailed information about claims or arguments made.
- Relevance: ensuring the information given in any argument is related to the matter at hand.
- Depth: considering all aspects of the issue.
- Breadth: taking into account alternative viewpoints or explanations.
- Logic: ensuring different parts of an argument flow logically from one to another.
- Significance: considering the important implications of one's ideas, focusing on the important elements.
- Fairness: making judgements that are objective and free from personal bias or conflicts of interest.

Based on these intellectual standards, five key CT skills are proposed. These are:

- Asking clear and precise questions.

- Collecting and assessing relevant information and interpreting it effectively.
- Drawing reasonable and logical conclusions/solutions.
- Considering alternative explanations, identifying, and evaluating assumptions.
- Communicating effectively with others to collaboratively devise solutions to complex problems.

Facione's model of CT skills

Using the Delphi method, Facione (1990) found a consensus among a panel of 46 scholars from various disciplines that CT involves a set of six cognitive skills. These are:

- Interpretation: understanding and explaining the meaning of information or data.
- Analysis: breaking down complex ideas into smaller components to understand their underlying structures.
- Evaluation: assessing the credibility, relevance, and significance of information or arguments based on the sources and the strength of the evidence.
- Inference: drawing logical conclusions and making predictions based on available information.
- Explanation: clearly articulating the reasons, assumptions, and rationale behind decisions or views.
- Self-regulation: examining one's thought processes, biases, and assumptions.

While there are conceptual variations in the literature on what CT is, some common features are evident. Three skills are common across all the above models: analysis, evaluation, and inference. Moreover, CT is seen as a composite cognitive skill comprising several sub-skills. These sub-skills are independent of each other but closely interrelated (Bailin, Case, Coombs, & Daniels, 1999; Dhakal et al., 2023). For instance, to draw a rigorous inference, it is necessary to first interpret the existing information and context, identify hidden assumptions, and evaluate the credibility of the information. In line with widely used CT frameworks, this thesis considers CT as a composite of a number of sub-skills.

Critical thinking dispositions

In addition to the skills dimension, CT is also defined as a set of affective dispositions. These dispositions are intellectual attributes referred to as the internal inclination, tendency, and willingness to think critically (Ennis, 2011; Facione, et al., 1995; Norris, 2003). These attributes include open-mindedness, truth-seeking, and inquisitiveness (see Bailin et al., 1999; Ennis, 1985; Facione, 1990; Halpern, 1998).

Ennis's model of CT dispositions

Ennis (2015) describes several key attributes of critical thinkers. First, they exhibit a disposition towards clarity and precision in articulating questions and presenting their reasoning. This inclination ensures that the foundations of their arguments are solid, fostering a dialogue rooted in coherence and logical progression. Moreover, Ennis (2015) considers critical thinkers to actively engage in the pursuit of knowledge, demonstrating a commitment to being well-informed. Recognising the importance of using reliable evidence to support claims and arguments, they seek out credible sources and observations. Contextual factors are also considered accurately, precisely, and comprehensively.

Thirdly, critical thinkers are believed to embrace the diversity of perspectives. They remain open-minded, receptive to alternative viewpoints, and steadfast in their willingness to entertain ideas that challenge their own preconceptions. They adopt a proactive stance towards adjusting their viewpoints when sufficient evidence is presented. Lastly, critical thinkers are willing to employ their CT skills. This is notable as CT skills and dispositions are different. For instance, a licensed driver might not be willing to drive a car. Similarly, individuals skilled in CT may not possess a positive CT propensity (Facione, Facione, & Giancarlo, 2000). Thus, according to Ennis (2015), the willingness to apply CT skills is considered one of the important conditions for critical thinkers.

Paul and Elder's model of CT dispositions

According to Paul and Elder (2012), critical thinkers possess seven essential traits. These are:

- Intellectual humility: acknowledging one's limitations and ignorance.
- Intellectual courage: having the courage to express what one believes is right, even if their views are not shared by the majority.
- Intellectual empathy: the ability to understand and appreciate the opinions and experiences of others.
- Intellectual integrity: being honest, fair, and consistent in presenting one's ideas, acknowledging sources, and being transparent about one's own biases and limitations.
- Intellectual perseverance: persistence in pursuing answers, not giving up easily on difficult problems.
- Confidence in reason: placing trust in empirical evidence, facts, and data, and prioritising evidence-based thinking over superstition or unfounded beliefs.
- Intellectual autonomy: the ability to think independently and critically, not being influenced by social pressures, biases, or external authorities.

Facione's model of CT dispositions

The consensus among researchers in Facione's (1990) Delphi programme is that critical thinkers are disposed to be inquisitive, well-informed, and confident in their ability to reason and draw inferences. Another common disposition is open-mindedness, displaying a willingness to consider alternative viewpoints, which is similar to Ennis's (2015) descriptions of critical thinkers' virtues. Furthermore, critical thinkers are fair-minded, aware of their biases and prejudices, and willing to regularly reflect on their views. All these CT dispositions are recognised as essential in our daily lives.

The group of researchers also identified problem-solving as a CT disposition. Well-cultivated critical thinkers exhibit a capacity for articulating clear questions or concerns, thereby laying a solid foundation for inquiry. They approach complexity with a methodical and orderly mindset, constantly seeking relevant information and selecting useful data while focusing on key matters. They prioritise precision in their quest for information. Finally, they persistently address problems, viewing challenges as opportunities for growth and learning.

The Delphi report provides a detailed account of CT dispositions (19 dispositions in total), some of which are intertwined and could be classified more meaningfully. For this reason, Facione and Facione (1992) later improved their framework, reclassifying these 19 dispositions into seven virtues:

- Analyticity: the disposition to analyse, reason, and evaluate information or arguments systematically.
- Inquisitiveness: the disposition to be curious, interested, and engaged in the pursuit of knowledge and understanding.
- Systematicity: the disposition to solve problems and make decisions in a methodical and organised manner, considering multiple factors and perspectives.
- Open-mindedness: the disposition to approach ideas, arguments, and evidence with an open mind, without bias or prejudice.
- Truth-seeking: the disposition to actively seek the truth and pursue knowledge through evidence-based reasoning.
- Self-confidence: the disposition to have confidence in one's reasoning abilities and judgement.
- Maturity of judgement: the disposition to exercise discretion in decision-making, considering the consequences of one's actions.

As shown above, a consensus on the precise nature of CT dispositions has yet to be reached. Researchers may describe CT dispositions differently, but common characteristics include open-mindedness, intellectual curiosity, empathy, and integrity (Sosu, 2013). To reconcile these overlapping and interconnected tendencies, this thesis regards CT dispositions as a composite mindset comprising several sub-sets of propensities.

Critical thinking styles

In addition to cognitive abilities and affective dispositions (Black, 2007), another aspect of CT is thinking styles (Lamm & Irani, 2011). The cognitive style of CT delineates an individual's inclination towards a specific approach over others when processing information or critically evaluating a particular issue (Gorham, Lamm, & Rumble,

2014). According to Lamm and Irani (2011), CT is a distinctive and deliberate type of thinking that facilitates individuals to reason through complexity using established standards.

There are two styles of CT: engagement and information seeking (Lamm & Irani, 2011). **Engagers** are more inclined to construct meanings from their surroundings and interactive communication. They possess an awareness of their environment, allowing them to recognise when situations necessitate sound reasoning to address a problem (Akins et al., 2019). Individuals who tend to engage also value the opportunity for open discussion in group settings. They are proficient and confident in presenting ways and processes of solving problems and drawing conclusions.

Information seekers, on the other hand, are keen on acquiring information (Lamm & Irani, 2011). They often reflect on their personal experiences and recognise biases and prejudices. Furthermore, these individuals understand that problems can be intricate, and there is usually not a single solution available. Thus, they actively seek as much information as possible to enhance their knowledge and be open to divergent points of view (Gay, Terry, & Lamm, 2015).

An individual's CT style lies somewhere on a continuum from engagement to information seeking. People tend to fall somewhere in between on this spectrum (Akins et al., 2019), meaning they exhibit a mixture of these two styles. As stressed by Lamm and Irani (2011), both styles are necessary, and ideal critical thinkers are expected to flexibly apply them in different contexts.

In this thesis, I define CT according to these three dimensions: cognitive skills, affective dispositions, and thinking styles (Baker, Lu, & Lamm, 2021; Ku, 2009).

2.2 Can critical thinking be taught?

The second debate concerns the teachability of CT. Willingham (2008), for example, recognised that despite efforts and different pedagogical approaches, substantial improvements in CT skills are rare and often not sustained over time. He then questioned the teachability of CT. From the perspective of cognitive science, the answer

is “not really” (p.21). One important aspect of CT is dispositions such as open-mindedness and inquisitiveness, which are relatively stable and inherent (Bailin et al., 1999; Ennis, 1985; Facione, 1990). This suggests that these traits are not easily modified through instructional interventions.

Moreover, as Paul and Binker (1990) suggested, effective CT teaching requires a willingness to question one’s own beliefs and assumptions. This idea is based on the assumption that all humans are innately egocentric and sociocentric. Egocentrism means that people pursue selfish ends and fail to consider the rights and needs of others, while sociocentrism implies that perceptions of the world are inherently shaped by biased perspectives centred around groups or collective identities (Paul & Elder, 2012). If people are unaware of these natural thinking obstacles, they are unlikely to benefit from CT instruction. Thus, attempts to teach CT may face resistance if people are not reflective and self-critical.

Additionally, there may be cultural boundaries to CT teaching. Some researchers believe that CT is a distinct Western cultural construct (Atkinson, 1997, 1998; Ramanathan & Kaplan, 1996). Atkinson (1997) considered CT as common sense shared by Westerners, as they are immersed in CT in everyday life. Fox (1994) and Ramanathan and Kaplan (1996) also argued that the Western norm of good writing demonstrates CT competence, whereas this is not a consensus in other cultures. Therefore, it may not be possible to teach CT across different cultures. For instance, cultures that emphasise collective harmony over individual critique may not value or foster the same type of CT promoted in Western education.

There is some empirical evidence offering mixed results regarding the effectiveness of CT teaching. For example, Behar-Horenstein and Niu (2011) reviewed several specific instructional methods of CT instruction in higher education, such as concept mapping, problem-based learning, and inquiry-based learning, but found no consistent results on the effectiveness of any instructional approach. This could be attributed to the different implementations of interventions. Even if the same instructional approach was adopted, it might have different impacts on CT skills. Similarly, Abrami et al.’s (2008) meta-analysis of 117 empirical studies suggests that it is feasible to teach CT, but the overall

impact is modest. Their subsequent review indicates that there are effective methods for teaching CT regardless of educational levels and disciplinary fields (Abrami et al., 2015). To gain a clear understanding of these interventions, CT instructional approaches will be discussed in detail below.

2.3 How can critical thinking be taught?

If CT can be taught, the questions that follow are: How can it be taught, and are there effective ways of teaching it? There are two main views on this: the domain-specific approach and the general approach. Some scholars maintain that CT is a general skill and should be taught across disciplines (Hare, 1999). This means that CT can be transferred to different domains, and its cultivation does not require students to have a deep understanding of specific disciplinary knowledge. Accordingly, a general approach to fostering CT is suggested (Davies, 2013; Van Gelder, 2005).

On the other hand, others suggest that CT is domain-specific and therefore requires background knowledge of the subject (Bailin et al., 1999; McPeck, 1981). For example, Willingham (2019) argued that each discipline has its fundamental logic, and subjects such as science and history may have different interpretations of the meaning of “knowing” something. Nilson (2021) also contends that it is not useful to have an overarching principle of CT due to the variations in terminology, context, and evaluation methods of CT outcomes across different academic disciplines.

A third approach combines both the domain-specific and general approaches. Some commentators argue that both approaches are necessary to foster CT (e.g. Ten Dam & Volman, 2004; Willingham, 2008).

To this end, Ennis (1989) proposed four methods of CT teaching: general, infusion, immersion, and mixed. Table 2.1 provides the classification and distinction of these four CT teaching methods.

Table 2.1 Typology of Ennis's (1989) critical thinking teaching methods

Approaches	An explicit objective	Uses subject-matter content
General	Yes	No
Infusion	Yes	Yes
Immersion	No	Yes
Mixed	General + infusion OR general + immersion	

General approach

In the general course, CT is taught independently of specific subject matter, and the CT objectives are made explicit to students. The general method of CT teaching has been widely adopted, and there is tentative empirical evidence on its effectiveness (Abrami et al., 2008; Behar-Horenstein & Niu, 2011; Marin & Halpern, 2011). Rimiene (2002), for instance, conducted a quasi-experimental study on the effectiveness of the general CT teaching approach on university students' CT skills. A separate CT programme was designed and implemented, comprising various activities such as reflective writing, active listening, and cooperative learning. Students participating in this CT programme acquired an understanding of CT principles, stages, and standards of sound reasoning, enabling them to effectively address problems. Experimental students outperformed their counterparts in all subsets of CT skills (i.e. analysis, evaluation, inference, deductive reasoning, and inductive reasoning), which indicated the positive impact of the general method. However, it should be noted that although 227 students participated in the study, the number of control students ($n = 150$) was much larger than that of experimental students ($n = 77$). This imbalance in student numbers may skew the results.

Marin and Halpern (2011) compared the impact of the general method and the embedded method of CT teaching on high school students' CT skills. The stand-alone course in this study was a web-based CT workshop that included analysing arguments, distinguishing between causation and correlation, identifying stereotypes, and making reasonable decisions. CT was also embedded in an introductory psychology course where students had opportunities to exercise CT skills, including analysing, interpreting, identifying logical relationships, and solving problems. There was also a wait-list group ($n = 24$), which formed the control. Both groups of students in the general and the

embedded workshops exhibited progress in CT skills. The former group (n = 28) made greater progress in CT skills than those in the embedded workshops (n = 16). However, the results should be treated with caution as there was no post-test CT score for the control group, and the study was limited to one school, with only 68 students in total. Moreover, the measurement of CT skills remained questionable. While the use of the Halpern Critical Thinking Assessment was appropriate, the absence of questions pertaining to the skill of comprehending likelihood and uncertainty poses a concern. The authors clarified that this skill was not covered in the instruction, so these questions were excluded. This suggests an element of teaching to the test.

El Soufi (2019) evaluated the general method of CT teaching on the CT skills of higher education students in Lebanon. She designed a general CT programme for students who learned English as a foreign language, covering content on common logical fallacies, correlation and causation, stereotypes, judging the credibility of sources, and making counterarguments. This was a cluster randomised trial conducted over two academic terms. She found that the general method had a positive impact on these students' CT skills (effect size = 0.30). The large sample size in each cell (198 in the experimental and 185 in the control groups) and the sensitivity analysis suggest the reliability of the positive impact. Notably, as acknowledged by the author, her role as the lecturer might be a major limitation. She was not only the researcher of the whole study but also responsible for the design and training of the intervention. She also delivered the intervention to six of the experimental classes. As she developed the test, designed, and delivered the intervention, there is a risk of unconscious bias or researcher expectation.

Infusion approach

In the infusion course, specific curriculum content is taught, and the development of CT is an explicit goal for students. The infusion approach to CT teaching has been primarily implemented in Kindergarten through 12th grade (K-12) contexts (Ventura, Lai, & DiCerbo, 2017). Perhaps this is because it does not contradict other educational objectives and is more easily accepted by schools as it is not an add-on to an already crowded curriculum (Zohar & Tamir, 1993). Additionally, the infusion approach can be easily integrated into many disciplines (Bensley & Spero, 2014; Zulkpli, Abdullah, Kohar, & Ibrahim, 2017).

Some research has shown that the infusion approach can increase students' CT skills (e.g. Bağ & Gürsoy, 2021; Zohar & Tamir, 1993; Zohar, Weinberger, & Tamir, 1994). For example, Zohar and Tamir (1993) demonstrated that the infusion approach in the Biology Critical Thinking Project could enhance the CT skills of ninth-grade biology students in Israel. As this was a pilot study including only 77 students in total, Zohar et al. (1994) expanded the intervention to 678 seventh-grade students. The results still showed that those who received the infusion CT biology curriculum made greater gains in CT skills than those in the business-as-usual group.

The infusion method can also be adopted in the teaching of the English language. Bağ and Gürsoy (2021) devised a CT-embedded English course and conducted a quasi-experimental study to evaluate its effectiveness. Their findings demonstrated the beneficial effects of the infusion method on the CT skills of seventh-grade students in Turkey. However, the result is tentative due to the small sample size (31 per cell). In addition, it is unknown whether the instructor taught both experimental and control classes. If this is the case, there might be a diffusion problem where the teacher unconsciously taught CT-relevant content to the control class. If there were two teachers, any impact might be due to teacher differences and cannot be exclusively attributed to the intervention.

Lin (2014) conducted a case study on the implementation of the infusion method in an English writing course at a Chinese public high school. This study indicated that the infusion method enhanced students' CT dispositions, skills, and language learning. However, the evidence is not strong as there were issues of contamination since the same teacher taught both the infusion and traditional English courses. Furthermore, the case study was based on only one case (Gorard, 2013) with no appropriate comparator.

While most studies reported positive effects of the infusion method (Ventura et al., 2017; Zulkpli et al., 2017), a few studies did not find similar positive results (e.g. Toy & Ok, 2012). This could be the result of publication bias, where studies with promising and positive results are more likely to be published (Song, Hooper, & Loke, 2013). Toy and Ok (2012), for example, tested the effectiveness of the infusion method in a vocational

pre-service teacher education programme in Turkey. They found that both treatment and control groups made similar improvements in CT dispositions over the course of a semester, suggesting no particular benefit of the infusion of CT compared to no treatment.

Immersion approach

Another approach to teaching CT is the immersion method. It is similar to the infusion method, but the cultivation of CT is not made explicit to students. A large proportion of studies evaluating CT teaching in higher education have adopted the immersion approach (Behar-Horenstein & Niu, 2011; Puig, Blanco-Anaya, Bargiela, & Crujeiras-Pérez, 2019; Tiruneh, De Cock, & Elen, 2018). For instance, in Tiruneh et al.'s (2018) systematic review of CT instruction in higher education, almost half of the 33 empirical studies employed the immersion approach. However, the immersion approach was considered the least effective compared to the other three approaches (Abrami et al., 2008; Al-Ghadouni, 2021; Tiruneh, Verburch, & Elen, 2014).

The immersion approach is more commonly used in higher education, perhaps because it can be embedded into domain-specific fields rather than as a stand-alone course (Puig et al., 2019). Additionally, the immersion approach requires the least resources and effort. In other words, it can be easily integrated into a course programme using the same course materials for discussions, debates, and other collaborative activities without making CT principles and procedures explicit (see Kamin, O'Sullivan, & Deterding, 2002; Semerci, 2006; Şendağ & Odabaşı, 2009; Yuan et al., 2008).

As an example, problem-based learning was employed in an undergraduate nursing course in China (Yuan et al., 2008). Students who participated in the course worked collaboratively to figure out solutions. Forty-six students were randomly assigned to either the immersion course or the usual lecture module. Results of this study showed that the immersion group outperformed the lecture group, which could support the usefulness of this approach. However, the result is tentative due to the small number of students, all from the same university.

Mixed approach

Some researchers have combined the general approach with either the infusion or immersion approach. Students taught using the mixed method engage in CT instruction tailored to their respective domains, supplemented by a stand-alone course that focuses on teaching the general principles of CT. Some reviews claim that the mixed approach is the most effective (Abrami et al., 2008; Abrami et al., 2015; Al-Ghadouni, 2021; Tilbury, Osmond & Scott, 2010). For example, Abrami et al.'s (2008) meta-analysis, which included 117 primary studies with 20,698 participants across all phases of education from K-12 and higher education, as well as adult learners in non-formal educational settings, found the mixed instructional approach to be the most effective and the immersion method the least promising. Examining the actual implementation of CT instructional interventions separately, their analysis revealed that pedagogical interventions that include explicit teaching produced the strongest effects. Studies where CT is merely mentioned within the curriculum description or outlined as a course objective, but with no explicit training of teachers, yielded the smallest effects. This highlights the necessity of teacher training on CT instruction.

Nevertheless, due to the heterogeneity across the included studies, it is difficult to determine if the stronger effects were driven by studies for a particular phase of education or by studies with weaker designs (e.g. with no comparators or small sample sizes) but reporting large effects. Meta-analyses and most prior reviews do not take these factors into account, instead lumping all effect sizes together and averaging them.

Findings can vary depending on how the intervention is implemented (Behar-Horenstein & Niu, 2011). For instance, Mahmood (2017) found no beneficial effect of the mixed method (explicit and embedded) approach for students in an initial teacher education programme in Pakistan. This was likely due to the short duration of the intervention, which lasted only four weeks and was delivered by one teacher. The process evaluation suggested there were problems with the preparation and training of the students.

In summary, there is a large body of work examining the effects of CT instruction, and the results are inconclusive. Given that the duration of the intervention varies from

several weeks (Mahmood, 2017) to a few months (El Soufi, 2019; Lin, 2014), and across a wide age range from primary-aged children to secondary and higher education students, this is not surprising. Nevertheless, previous reviews (Abrami et al., 2015; El Soufi & See, 2019) suggest that the explicit method is most promising, and identify classroom dialogue, the use of authentic or situated problems or examples and mentorship as specific strategies that are helpful in developing CT skills.

In the context of China, where schools adhere closely to prescribed syllabi and textbooks set by the Department of Education, the infusion approach seems to be the most pragmatic method for integrating CT into the regular curriculum. The infusion approach appears to be a feasible strategy, which is more likely to be accepted by schools in China.

Additionally, the most recent English curriculum plan and standards for primary and secondary schools in China have clearly outlined the desired thinking outcomes for students, involving interpretation, analysis, evaluation, inference, explanation, deduction, assumption, and self-regulation (MoE, 2022b). The integration of the infusion approach within the English curriculum at secondary schools is particularly advantageous due to the mandatory nature of the English language subject.

The infusion method is also widely used in K-12 education (Ventura et al., 2017), with secondary-aged students (11-15 years old) demonstrating greater receptivity to CT instruction compared to postsecondary learners (Abrami et al., 2008). With these considerations in mind, the primary research for my thesis will focus on the infusion approach to teaching CT in secondary schools in China.

2.4 How is critical thinking measured?

To determine whether an approach is effective in fostering CT, we need to understand how CT is assessed. As discussed in Section 2.1, there are three components of CT: skills, dispositions, and styles. A number of assessment tools have been developed to measure these components.

Measuring CT skills

The California Critical Thinking Skills Test (CCTST) is one of the most widely used tests of CT skills. It uses exclusively multiple-choice questions and is often applied in higher education settings to examine students' readiness for academic and professional success (e.g. Bycio & Allen, 2009; Din, 2020; Jacob, 2012). Two other popular standardised CT tests are the Watson-Glaser Critical Thinking Appraisal (WGCTA; Watson & Glaser, 2012) and the Cornell Critical Thinking Test (CCTT; Ennis, Millman, & Tomko, 2005a). They assess different sub-sets of CT skills. The WGCTA aims to assess inferences, recognition of assumptions, deduction, interpretation, and evaluation of arguments. There are two versions of the Cornell test: the CCTT-Level X for students in Grades 5-12 and the CCTT-Level Z for students at a higher academic level or in higher education institutions. The former evaluates skills of induction, deduction, credibility of sources, and identification of assumptions, whereas the latter includes three additional abilities: semantics, definition, and prediction in planning experiments (Ennis, Millman, & Tomko, 2005b).

These three standardised tests are commonly used for evaluating CT skills. All of them use multiple-choice questions, which are cost-effective, easy to administer, and objective (Ventura et al., 2017). However, some critics argue that a singular format may not provide a comprehensive assessment of CT skills (Halpern, 2005; Ku, 2009; Rear, 2019). The Halpern Critical Thinking Assessment (HCTA; Halpern, 2005) was developed to address this issue. The HCTA assesses students' verbal reasoning, argument analysis, hypothesis testing, handling likelihood and uncertainty, and decision-making and problem-solving through real-world scenarios. By incorporating both multiple-choice and constructed-response items, the HCTA allows for a more thorough exploration of students' cognitive processes and reasoning capacities. However, it is less widely adopted than the earlier three tests as it takes longer to complete and marking is more subjective and time-consuming, especially when there are discrepancies among raters. Concerns have been raised about the reliability of such assessments compared to the more objective multiple-choice type questions (Liu, Frankel, & Roohr, 2014; Lee, Liu, & Linn, 2011).

Despite their differences, there are similarities across all these different CT assessments. They all assume that CT skills can be divided into discrete and measurable sub-skills (Rear, 2019), although there is little evidence that these sub-skills exist independently of each other in real-world contexts (Lai, 2011; Liu et al., 2014; Hassan & Madhum, 2007). Bernard et al. (2008), for example, show that sub-skills such as drawing inferences, identifying assumptions, making deductions, and evaluating arguments are interconnected. However, educators believe that there is value in measuring specific CT skills, as it allows for assessing CT skills across various domains, thus providing richer diagnostic insights (Ventura et al., 2017). This pragmatic approach may explain why much of the current research continues to adopt these established assessments to evaluate the effectiveness of interventions on CT skills.

Measuring CT dispositions

There are different methods for assessing CT dispositions, including self-report surveys, interviews, situational judgement tests, and performance tasks. Self-report surveys measure self-perceptions and preferences, but they may not reflect actual behaviours. Interviews can elicit examples and evidence of CT dispositions, but they may not capture the complexity or diversity of situations. Situational judgement tests simulate realistic scenarios and dilemmas, but they cannot measure an individual's motivations. Many of these tests have not been validated or standardised at scale, and their suitability for students of different age groups is not established. Although some instruments, such as the Need for Cognition Scale and its Five-Factor Inventory (e.g. Cacioppo, Petty, Feinstein, & Jarvis, 1996; Costa & McCrae, 1992), are used by researchers to measure CT dispositions, they are designed to measure general thinking rather than specific CT dispositions.

Currently, the only standardised and validated instrument specifically developed to measure CT dispositions is the California Critical Thinking Disposition Inventory (CCTDI). This tool is suitable for undergraduate and graduate students. It was designed by Facione and Facione (1992) to measure seven key dispositions of CT: analyticity, inquisitiveness, systematicity, open-mindedness, truth-seeking, self-confidence, and maturity.

Several studies have investigated the reliability and validity of the CCTDI, but there are questions about its sub-scales. For instance, in the initial proposal of the CCTDI, Facione and Facione (1992) conducted a pilot study involving 156 participants who were high school students, undergraduates, and post-baccalaureates. The overall alpha coefficient for the instrument was 0.91, with sub-scale coefficients ranging from 0.71 to 0.80, indicating that the items are internally consistent (i.e. they measure the same construct). However, when the instrument was administered to 499 undergraduates from different disciplines such as history, nursing, and education, the sub-scale coefficients ranged from 0.57 to 0.78, indicating significant variability (Walsh & Hardy, 1997).

A recent meta-analysis incorporating 87 studies reports an overall alpha value of 0.83 for the CCTDI (Orhan, 2022). The alpha coefficients for its sub-scales ranged from 0.56 to 0.74. This meta-analysis also suggests that the alpha value tends to be higher when the CCTDI is administered to university students. These findings highlight the variability in reliability estimates across different populations and point out the importance of considering the context in which the CCTDI is employed.

Some studies suggest that the variability in reliability and validity of the CCTDI may be attributed to issues such as excessive item loading (Bondy, Koenigseder, Ishee, & Williams, 2001) and construct overlap (Liu & Pásztor, 2022). In an attempt to address these challenges, some researchers have revised and shortened the CCTDI (Liu & Pásztor, 2022; Walsh & Hardy, 1997). Yoon (2004), for instance, recommended reducing the original to 27 items and focusing on the key elements of intellectual curiosity, systematicity, prudence, objectivity, self-confidence, healthy scepticism, and intellectual fairness. Similarly, Sosu (2013) revised the CCTDI into a shorter scale emphasising critical openness and reflective scepticism. However, the validation of the new version remains questionable, as it has primarily been tested on students from the same programme. This raises uncertainties about their generalisability to a broader population.

Quinn, Hogan, Dwyer, Finn and Fogarty (2020) proposed a model with six distinct subscales for measuring CT dispositions: open-mindedness, reflection, attentiveness,

organisation, intrinsic goal motivation, and perseverance. This model was developed through consultation with both educators and students, aiming to comprehensively capture various facets of CT dispositions. Building on this framework, Liu and Pásztor (2022) introduced an innovative instrument that focuses on three key components: instant judgement, self-efficacy, and habitual truth-digging. They contend that this approach can address concerns related to overlapping constructs in previous instruments. The development of these refined measurement tools reflects a concerted effort to enhance the precision and applicability of assessments for CT dispositions.

Measuring CT styles

CT styles reflect the approaches that individuals employ in their reasoning and problem-solving endeavours (Lamm & Irani, 2011). Based on the conceptualisation of CT styles, Lamm and Irani (2011) developed The University of Florida Critical Thinking Inventory (UFCTI) to assess two CT styles along a continuum, ranging from engagement to information seeking. The UFCTI is not a measure of whether an individual is either a good or poor critical thinker. Its purpose is to identify how individuals learn and think (Lamm, 2015).

The UFCTI is a self-report questionnaire consisting of 20 items, with 13 questions on information seeking and 7 on engagement. Each item is scored from 1 (strongly disagree) to 5 (strongly agree). Scores for the two constructs are calculated separately and then weighted and added together to form the total score. Respondents who achieve a total score of 73 or above are identified as “Seekers”, while those scoring 72 or below are categorised as “Engagers”. The ideal score is somewhere in the middle, demonstrating that the person is balanced in the way they learn and think.

The UFCTI is reported to be a reliable measure after being tested rigorously on multiple populations (Lamm & Irani, 2011). This suggests that the tool is highly likely to yield similar outcomes under consistent conditions. This instrument has also been translated into Chinese to evaluate the validity and reliability of the instrument for Chinese international students (Baker et al., 2021). In a study involving 148 undergraduate agricultural students in China, the questionnaire was completed online, and confirmatory factor analysis was conducted to establish the equivalence between the

English and Chinese versions of the UFCTI (Baker et al., 2021). The results showed that the Chinese version had high validity and reliability, measuring the same underlying CT style as the original English version. Internal reliability was 0.84 and 0.92 for engagement and information seeking, respectively. All of the factor loadings were higher than 0.5, demonstrating an adequate level of construct validity.

2.5 Chapter summary

This chapter discusses the common debates around CT: what it is, how it is measured, and whether it can be taught. Although there are many interpretations and definitions of CT, it generally involves a combination of skills, dispositions, and styles. The infusion method of CT appears to be the most relevant to the premise of this PhD research as it is the most promising and can be easily embedded into the existing school curriculum without the need for additional add-on lessons. Thus, it is less likely to meet with resistance from schools in China, which are very exam- and textbook-oriented, as the infusion method can use the teaching resources and textbooks already used in schools. This justifies the use of the infusion method as the intervention in my primary research. The discussion on the different CT assessment tools also provides justification for the assessment tool I used in this research.

Chapter 3 The critical thinking of Chinese students

This chapter first introduces how Chinese students' critical thinking (CT) is portrayed in the literature and then presents how existing research addresses this issue. Through a critical review of previous studies, this chapter demonstrates the importance of conducting a systematic review of Chinese students' CT.

3.1 Common perceptions of Chinese students' critical thinking

With the increasing number of Chinese students studying in Western universities over the last decade, there is a growing interest in their learning skills and dispositions. A common stereotypical perception is that Chinese students are deficient in CT (Song, 2014; Xu, 2021). For example, Lucas (2019) suggested that Chinese learners lack training in CT skills such as analysing and evaluating information. They have also been reported to face challenges in articulating their ideas in international class discussions (Guo & O'Sullivan, 2012). Their learning has been described as superficial (Watkins & Biggs, 1996), focusing on memorisation rather than interpretation or analysis. They struggle with analysing information, searching for credible sources, questioning assumptions, evaluating arguments, and constructing their own viewpoints (Turner, 2006). They acknowledge a lack of understanding of how to apply CT in their learning (Zhong & Cheng, 2021).

Influenced by Confucian culture, Chinese students have traditionally been educated to show great respect for knowledge from teachers, preferring a teacher-centred style in class (Kirkbride, Tang, & Westwood, 1991). This has made them passive recipients of knowledge. They are observed to be silent in class, rarely sharing their views or actively engaging in discussions (Ping, 2010). This also influences their epistemic beliefs. Those who believe knowledge is certain tend to be less open-minded (Chan, Ho, & Ku, 2011), exhibiting less willingness to consider alternative viewpoints.

Some studies using standardised measures to assess Chinese international students' CT appear to confirm this assumption. Lun, Fischer, and Ward (2010), for instance, used the Halpern Critical Thinking Assessment Using Everyday Situations to evaluate the CT skills of 24 Chinese students studying at a university in New Zealand. The standardised test score for Chinese students was -1.26 (SD = 1.70), indicating low

proficiency in CT skills. It should be noted that this small group of Chinese students, already studying overseas, may differ from those in local Chinese institutions in terms of demographic information such as socioeconomic status (SES) and academic backgrounds. Another issue is that these Chinese participants were tested in English, which was their second language. Thus, the conclusions should be treated with caution, as English language proficiency may be mistakenly used to measure CT (Moosavi, 2021).

CT in local Chinese students has also been investigated, and it seems to support the claim. For example, Ip et al. (2000) tested the CT dispositions of 122 Chinese nursing students at a university in Hong Kong. Their results indicated that Chinese students showed a negative disposition towards CT, with truth-seeking being the lowest. However, this cannot be generalised to the whole Chinese population due to the small, selective sample limited to one university's nursing programme.

The study by Zhang and Lambert (2008) also appeared consistent with the stereotypical perception. They used the California Critical Thinking Disposition Inventory (CCTDI) to assess the CT dispositions of 100 university students in a nursing programme in central China. Based on the lower average total CCTDI score, they concluded that these students did not have positive dispositions towards CT. However, due to the selective sample restricted to one discipline, it is far from conclusive to state that Chinese students are poor at CT. Moreover, while the authors claimed that the scores of Chinese students' CT dispositions were lower than those from Western cultures, the comparison was not rigorous. They cherry-picked two studies that reported higher CT dispositions among Western students compared to Chinese students. It is difficult to conduct a fair comparison when the sample size, participant characteristics (e.g. age, gender, and disciplines), and measurement of CT dispositions differ across studies.

The stereotypical perception of Chinese students as passive learners lacking criticality could have a damaging and self-fulfilling effect on students. Some scholars, such as Song (2014) and Xu (2021), argue that perpetuating such negative perceptions can lead Chinese students to internalise a discourse portraying their CT competency as deficient. Thus, these students may exhibit diminished confidence and reticence in expressing

their viewpoints, thereby inadvertently reinforcing Western academics' perception of them as lacking critical awareness (Li, Chen, & Duanmu, 2010).

3.2 Possible explanations for the stereotype

The stereotypical image of Chinese students as passive and uncritical consumers of information may be attributed to several factors. First, it could be due to the ambiguity in the definition of CT (Guo & O'Sullivan, 2012; Lucas, 2019). As discussed in Chapter 2, one major debate concerning CT is its conceptualisation. The diverse definitions of CT are reflected in the multiple ways Chinese students understand the concept. In Huang's (2008) study, many Chinese students openly acknowledged that they had no idea what the concept entailed. Even worse, the word "critical" conveys implicit negativity in both English and Chinese (O'Sullivan & Guo, 2010; Wu, 2011). Some Chinese students equated CT with negative thinking (Guo & O'Sullivan, 2012), focusing exclusively on opposing positions and identifying the disadvantages of arguments (Fakunle, Allison, & Fordyce, 2016).

Another potential explanation is a cultural one (Atkinson, 1997). Some scholars trace CT back to the age of Socrates (Facione et al., 1995) and claim that CT is a distinct and unique product of Western culture, incompatible with Asian culture (Atkinson, 1997). This conceptualisation implies that Chinese students naturally lack CT. Guo and O'Sullivan (2012) note that the Chinese culture of conformity, respect, and reverence for authority might explain Chinese students' reluctance to question and argue. The Chinese preference for the middle way differs from the Western preference for independent thought, reason, and the ability to debate and argue publicly (Durkin, 2011). Influenced by this culture, Chinese students might experience considerable discomfort when they first encounter Western teaching styles (Lucas, 2019).

An alternative explanation for the perceived lack of CT in Chinese students could be their lack of confidence and proficiency in English. In Western English-speaking universities, English is the language used in tasks involving CT. Constructing a coherent argument requires a high level of language proficiency. The difficulties Chinese students face in expressing themselves could be misconstrued as a lack of engagement in CT or a deficiency in cognitive skills. Western academics often interpret

the perceived reluctance of Asian students to participate in class discussions as a lack of CT (e.g. Durkin, 2008; Lee & Carrasquillo, 2006). Previous research has shown a positive correlation between language proficiency and performance in CT skills tests (e.g. Clifford, Boufal, & Kurtz, 2004; Taube, 1997). When Asian students are tested in their first language, they perform well and sometimes better than students in Anglophone countries (OECD, 2014). Floyd (2011) also supports the importance of language in CT tests. When Chinese speakers were tested with the Watson-Glasser Critical Thinking Appraisal in their native language, they performed better than when tested in English.

It has to be noted that when Chinese students' CT has been assessed in Western contexts, it is often based on the original English versions of test instruments, which were almost all developed by Western researchers (e.g. WGCTA, CCTT, CCTDI, and UFCTI). As Qasserras and Qasserras (2023) argue, while CT, defined as critical intelligence, is not culture-specific, language proficiency plays an important role. Naturally, it would be harder for Chinese students to take CT tests in English than in their native language (Floyd, 2011). Therefore, requiring international students to take CT tests in a foreign language adds to their cognitive load, impacting their performance (Moosavi, 2021; Qasserras & Qasserras, 2023). However, the language factor is not always considered when assessing CT in non-native English speakers (e.g. Lun et al., 2010).

The lack of understanding of Chinese students' educational experiences by Western academics could also explain the stereotypical perceptions. Chinese students are often influenced by their prior educational experiences (Zhang, 2017), but Western academics tend to view students of other nationalities through their own cultural lens (Paton, 2005; Turner, 2006). For those studying in the UK for the first time, unfamiliarity with Western academic traditions, such as academic writing that requires a high level of critical analysis and the ability to present opposing viewpoints, may be mistaken for a lack of CT (Turner, 2006). Paton (2005) pointed out that Chinese students' challenges in CT were due to insufficient knowledge and experience in a new situation. Chinese students may be accustomed to their role as passive recipients of knowledge in a traditional teacher-centred model of instruction (Lucas, 2019).

Zhang (2017) posits that the “Four Treasures” curriculum in Chinese higher education contributes to the perceived lack of CT among Chinese students. This curriculum includes four compulsory modules of party ideology propaganda, taken by all Chinese undergraduate students: “The Fundamentals of Marxism,” “Maoism and Chinese Characteristic Socialism,” “The Outline of Modern Chinese History,” and “Moral Thoughts, Legal and Civic Education.” These modules have been criticised for stifling CT. The curriculum presents content as absolute truth that cannot be questioned or challenged. Students are taught to defer to established sources of knowledge rather than engage in independent research and critical evaluation. Precise replication of Marxist knowledge is considered a criterion of excellence, and this regimented learning discourages students from challenging conventional wisdom, questioning authority, and exploring alternatives. Therefore, this limits opportunities for students to develop their CT skills (Zhang, 2017).

In summary, the stereotypical perceptions of Chinese students’ CT may stem from several sources — vague definitions of CT, cultural differences, English as a second language, and prior educational experiences.

However, there is no evidence that Chinese students are actually weaker than students of other nationalities. Yet, researchers and academics are quick to design curricula to address this perceived issue. Perhaps this is not a uniquely Chinese phenomenon (See, 2016). Students in Western democracies (e.g. the UK and the US) also show a lack of criticality. In a study involving 237 first-year undergraduates at two UK universities, See (2016) found that most students seemed to lack critical awareness when reading academic papers. They rarely questioned or challenged research findings, particularly when reading recently published or peer-reviewed articles. A report also indicated a lack of argumentative skills among British undergraduates (Independent, 2006), who seemed to lack the ability to present a reasoned argument or express themselves in writing.

Things may be no better in the US. In a study involving more than two thousand students, Arum and Roksa (2011) found that forty-five percent did not improve their CT skills in university. These students could not distinguish facts from opinions, present

clear and well-reasoned arguments, or synthesise and evaluate existing information. A decade later, a national survey involving 1,010 Americans investigated the state of CT (Reboot Foundation, 2021). The average age of respondents was 35, and 66% had a bachelor's degree or higher. According to this survey, while 95% of participants agreed on the importance of CT skills, 85% believed that the general public lacked CT skills.

3.3 Challenges to the common perceptions

Although Chinese students are commonly perceived as lacking in critical awareness, some scholars have begun to challenge this stereotypical image (Lu & Singh, 2017; Tian & Low, 2011; Xu, 2021). They highlight the necessity of understanding CT within the Chinese context (e.g. Heng, 2018; Lu & Singh, 2017), emphasising that the interpretation of CT in China should deviate from Western cultural traditions. It would be misleading to interpret Chinese students' CT solely through a Western lens. For example, Lu and Singh (2017) noted that Chinese students studying in Anglophone universities are multilingual, and their linguistic repertoire should be considered when evaluating their CT abilities. Heng (2018) further clarified that this divergence might not necessarily indicate a deficiency in CT among Chinese students. Instead, it may reflect different communication styles.

To develop a contextualised concept of CT in China, researchers have sought insights from Chinese school leaders (Tan, 2020), university lecturers (Zhang, Yuan, & He, 2020), and students (Chen, 2017). However, there has been no consensus. Tan (2020) asked three open-ended questions to 16 school leaders from Shanghai, inquiring about their understanding of CT, whether CT was promoted in their schools, examples of such promotion, and challenges faced in promoting CT. The findings indicated that CT was interpreted as personal inquiry and problem-solving, and its promotion was primarily concerned with the current education reform, high-stakes assessments, and common socio-cultural values. Meanwhile, university teachers who taught English and Foreign Languages were asked to articulate their understanding of CT via a questionnaire (Zhang et al., 2020). They defined CT predominantly through dispositions such as having multiple points of view, being fair-minded, open-minded, and truth-seeking, and skills such as making reasonable and logical judgments, analysing, and problem-solving. Additionally, 46 Chinese college students were interviewed about their

conceptualisation of CT (Chen, 2017). Most provided both definitions and examples of CT, highlighting cognitive skills, intellectual autonomy, and the consideration of pros and cons. These findings indicate the complexity of defining CT in China, as different people have different understandings of it.

Some empirical research findings also challenge the stereotypical claim. For instance, Li (2013) presented evidence from writing assignments of two high-achieving Chinese students studying at a university in Hong Kong to challenge the stereotype. The author indicated that both students demonstrated CT in their learning, such as clarifying views, presenting arguments logically, referring to relevant literature, listing sources of information, and integrating them meaningfully. However, this study is based on only two students, which cannot be representative of Chinese students in general. It could also be biased due to the author's potential selection of advantageous aspects of CT in the students' work.

A recent cross-sectional study used the California Critical Thinking Skills Test (CCTST) and CCTDI to evaluate CT skills and dispositions, respectively (Ng, Cheung, & Cheng, 2022). It involved 209 Chinese college students majoring in Science, Engineering, and Health Studies in Hong Kong. This group of Chinese students demonstrated a moderate level of CT skills and positive CT dispositions, including open-mindedness, analyticity, confidence in reasoning, and inquisitiveness. However, the findings cannot be generalised as they were limited to only one institution (i.e. Hong Kong Community College).

Eighteen Chinese undergraduates studying in the US were interviewed to share their experiences and opinions on CT (Heng, 2018). Over time, half of them indicated that they became more open to various opinions and constantly reminded themselves not to accept information without serious thought. While this opposes the deficit assumption of Chinese students' CT, it is based on subjective impressions, which may be notoriously unreliable. Additionally, the findings should be treated with caution, as this group of students may already come from better academic or family backgrounds than local Chinese students.

Perhaps the largest-scale cross-sectional study on the CT of Chinese students was conducted by Loyalka et al. (2021). They employed a Chinese version of the Critical Thinking Exam from the Educational Testing Service to assess the CT skills of 9,247 Chinese undergraduates. The study also included 17,455 Indian students, 4,703 Russian students, and 973 US students. Their results showed that first- and second-year Chinese students demonstrated similar levels of CT skills as US students, and they performed better than Indian and Russian students. However, by the fourth year, Chinese university students still demonstrated an advantage in CT skills over Indian students but not over their Russian and US counterparts. The results seem credible due to the large sample size and use of standardised CT tests. Notably, this study only focused on computer science and electrical engineering disciplines in higher education, so it remains unknown whether these conclusions apply to other disciplines and educational levels.

3.4 Strategies for improving critical thinking of Chinese students

Due to the stereotypical assumption that Chinese students lack CT, some researchers have sought ways to address this perceived shortcoming. Huang (2008), for example, suggested that Chinese students should acknowledge cultural and learning differences, read broadly and deeply, recognise different perspectives, and critically present their own arguments. These suggestions were based on interviews with five lecturers involved in the Tourism and Hospitality Management postgraduate programme. Similarly, Fakunle et al. (2016) proposed an introductory course, “Critical Thinking for Academic Purposes”, for postgraduate students, after interviewing six Chinese postgraduates majoring in Education in Scotland. Badger (2019) also suggested a tailored curriculum to address gaps in CT skills among Chinese international students. Based on the perceptions of 12 Chinese students and two university faculty members, the author claimed that the Intensive English Programme, which integrated creativity and CT skills, effectively cultivated international students’ analytical skills. Moreover, Zhong and Cheng (2021) recommended that asking open-ended questions and organising group discussions could facilitate the development of CT, based on interviews with 16 Chinese students enrolled in one-year Master’s programmes at a UK university.

The call to enhance Chinese students' CT has not only been investigated in international contexts but also in local Chinese environments. At Tongling University, three major teaching strategies—group discussions, concept mapping, and analytical questioning—were considered effective in facilitating Chinese students' CT skills (Wang & Seepho, 2017). This finding was based on a questionnaire where 50 students indicated their level of agreement on the helpfulness of these methods. Additionally, after analysing lesson transcripts in a class of 39 first-year undergraduates, dialogic instruction in an English reading course was considered effective in improving the CT of Chinese students (Cui & Teo, 2023).

While these studies attempt to address the perceived deficiency in Chinese students' CT, the effectiveness of these strategies should be questioned. First, the small sample sizes limit the representativeness of the findings. Most samples involved fewer than 50 participants, limiting the general applicability of the results. Secondly, there are no standardised tests on CT outcomes. These studies relied on subjective views, which are unlikely to yield reliable and valid results. Thirdly, the lack of comparisons is a major flaw. No pre- and post-tests were used to track changes in CT outcomes, and the effectiveness of interventions cannot be established without knowing the baseline or the situation of those who did not receive the intervention.

To address methodological weaknesses common in research on Chinese students' CT, some studies have adopted experimental designs, such as randomised controlled trials. For example, Tiwari, Lai, So, and Yuen (2006) evaluated the impact of problem-based learning (PBL) on Chinese students' CT. Using the CCTDI to measure students' CT dispositions, they found that after one academic year, the PBL group of 40 undergraduate nursing students at a university in Hong Kong exhibited greater progress in CT dispositions compared to the 39 students who attended usual lectures. Similarly, Yuan et al. (2008) tested the effect of PBL on Chinese students' CT skills, involving 46 undergraduate nursing students in China, with 23 in the experimental group. The study showed that PBL students made greater gains in CT skills, measured using the standardised CCTST. However, both studies were small-scale and conducted in single institutions, limiting their robustness and generalisability.

Moving the context from higher education to secondary schools, Hwang, Huang, Wang, and Zhu (2021) evaluated the effects of a concept mapping-based problem-posing approach in a Taiwanese secondary school. Two classes involving 40 students were recruited, with 21 in the experimental group and 19 in the control group, both taught by the same teacher. After a 150-minute intervention, experimental students showed an increase in CT tendency. However, the findings are tentative due to the small sample size, short duration of implementation, and potential bias from the same teacher instructing both groups.

Fung's (2014, 2017; Fung & Howe, 2014) studies seemed to overcome these shortcomings. They recruited 140 secondary students and four teachers from two Hong Kong schools to examine the effects of different pedagogies on students' CT dispositions and skills. Seventy students attended conventional classes, while the others joined either self-directed group work ($n = 40$) or teacher-supported group work ($n = 30$). After 10 hours of implementation, students in the teacher-supported group work showed the most progress in CT skills and dispositions. This study carefully considered research design and expanded participants across different schools, with an appropriate intervention duration, as CT cannot be developed overnight. However, a larger sample would make the findings more convincing.

Overall, while research has evaluated several approaches to enhance Chinese students' CT skills and dispositions, most studies are small-scale (involving fewer than 150 participants in total) and limited to one institution (e.g. Tiwari et al., 2006; Yuan et al., 2008), raising questions about generalisability. Additionally, in many cases, only a few teachers (sometimes the researchers themselves) deliver the intervention, and sometimes the same teacher is responsible for both experimental and control classes (e.g. Hwang et al., 2021). This could introduce biases where the teacher may consciously/unconsciously favour the treatment class or unconsciously expose control students to some elements of the intervention. When different teachers are allocated to teach separately (e.g. Fung, 2014, 2017), any differences in outcomes could be attributed to teacher differences. Therefore, it is important to involve more teachers across different schools. Finally, due to time and cost constraints, some strategies for CT improvement are not delivered for long periods (e.g. Hwang et al., 2021). The

fidelity of intervention has rarely been assessed. None of these studies have examined the long-term impact of these CT approaches. Given these concerns, although many studies have proposed “effective” approaches to improve Chinese students’ CT, their evidence is weak. To address the methodological weaknesses identified above, a rigorous randomised controlled trial was adopted in this thesis.

It should also be noted that all these studies start with the premise that Chinese students lack CT skills. No attempt was made to establish whether Chinese students are indeed lacking in CT. To do so would require comparative studies where Chinese students’ CT is compared with that of students of other nationalities. Researchers in this field may take it for granted and then try to find ways to improve this situation. However, efforts and resources might be wasted designing interventions to improve the CT of Chinese students if there is no evidence that they lack CT. We would be solving a problem that does not even exist in the first place. Hence, this thesis first examines the common assumption about the lack of criticality among Chinese students through a systematic review. It is unique in the following ways:

- Exclusively including validated standardised tests of CT: Some existing studies judge Chinese students’ CT based on subjective impressions. This is inconsistent, as different people hold different views on CT. These various understandings make it difficult to measure CT accurately. A more reliable evaluation of CT skills would be the use of standardised tests (Gorard et al., 2017). While some researchers may be concerned about the format of multiple-choice questions that may involve guessing (Snyder, Edwards, & Sanders, 2019), the pre-specified evaluation criteria and the validation of testing items allow for a high level of objectivity (Norris, 1989).
- Including comparisons with students from other nationalities: The assumption that Chinese students lack CT implies a comparison. Without comparisons with students from other nationalities, it is not possible to conclude whether Chinese students have higher, lower, or comparable CT skills to students in Western democracies (Gorard, 2013). What are we comparing Chinese students’ CT with, whose CT are we comparing, and what does the norm look like? Even if Chinese students show positive results towards CT, it remains unknown whether they

would perform better or worse in international comparisons. Most research into students' CT does not include a comparator, yet makes bold claims about the low levels of CT skills among Chinese students. This is absurd and widely accepted.

- Drawing more general conclusions: Research in this area has been restricted to higher education (e.g. Loyalka et al., 2021; Yeh & Chen, 2003) and the nursing discipline (e.g. Yuan et al., 2008; Zhang & Lambert, 2008). This makes it difficult to generalise findings to the whole Chinese population. A systematic search for available studies may uncover CT among Chinese students from different educational levels and various disciplines.
- Focusing exclusively on Chinese students and avoiding publication bias: There is a dearth of reviews investigating the CT of Chinese students (e.g. Huang, 2019; Tian & Low, 2011). Those that exist often focus on a broader group, such as Asian students in general (e.g. Indra, 2019; Salsali et al., 2013). For example, Salsali et al. (2013) compared the CT dispositions of Asian nursing students with those from other continents. Although they claimed to use a systematic method, there was no appraisal of the strength of evidence of the included studies, making it impossible to judge the evidence. Their review also included only peer-reviewed papers, introducing publication bias (Song et al., 2013), as studies reporting large, positive results are more likely to be published. These studies tend to be small-scale, using researcher-developed test instruments or lacking a comparator (i.e. single group, pre-post design). A large number of high-quality, large-scale, well-controlled studies remain unpublished. Cheung and Slavin's (2016) review found that 59% of these high-quality studies were unpublished. Excluding such studies can skew results and lead to misleading conclusions (Slavin & Neitzel, 2020; Slavin, 2020).
- Systematic searching and critical quality assessment: Among the few reviews focused solely on Chinese students, none were systematic or critical. Tian and Low (2011), for example, provided critical insights into studies about Chinese students' CT dispositions but did not search systematically, so no studies considering the CT skill dimension were found. Huang's (2019) review focused on high school students but failed to evaluate the trustworthiness of the evidence.

In other words, threats to validity, such as sampling strategy, sample size, attrition, and conflict of interest, were not considered.

For these reasons, my research reviews credible studies that compare Chinese students' CT with that of other nationalities, using validated standardised tests.

3.5 Chapter summary

This chapter illustrates that there is no consensus on Chinese students' CT. While some researchers take it for granted that Chinese students are poor at CT, others challenge this stereotypical assumption. It also justifies why a systematic review that compares the CT of Chinese students with other nationalities is necessary and unique. Evidence is expected to be established for the common assumption about the lack of criticality of Chinese students.

Chapter 4 Critical thinking education in China

Over the last two and a half decades, China has embarked on education reforms, spearheading bold and ambitious changes in the school curriculum. One significant shift is away from an emphasis on knowledge acquisition towards a focus on critical thinking (CT). However, despite these policies, teaching and learning in the classroom, for the most part, remain unchanged.

This chapter discusses the policy development of CT in China and explores why, despite these reforms and changes in the curricula, little progress has been made in classroom practices. One of the reasons is that teachers are not ready for change. They are ill-prepared for these reforms. This has implications for teacher training. Finally, this chapter explores some of the challenges in implementing the new CT education in China. Overall, this chapter clarifies the contexts of delivering CT teaching in China.

4.1 The development of policy on critical thinking education

Modern education reform in China can be traced back to the 1990s when significant changes were made to modernise the education system to align with the country's rapid economic and social growth. Revisions to the school curriculum were introduced as part of this reform (MoE, 2005), with a strong emphasis on Science, Technology, Engineering, and Maths (STEM) subjects, as well as practical skills relevant to a modern knowledge-based economy.

In line with these reforms, the Ministry of Education issued curriculum standards for primary, high school, and college students in 2005. This document provided detailed guidance for schools, emphasising the integration of modern teaching methods to foster CT, creativity, and problem-solving skills. The focus of teaching and learning shifted from the acquisition of basic knowledge and skills to a more holistic development approach (Paine & Fang, 2006; Ryan, Kang, Mitchell, & Erickson, 2009). A student-centred approach was advocated, and new assessments were planned to support students' personal and social development. Quality became a central focus of the educational reform, with an emphasis on teaching standards and enhancing teacher training programmes.

As part of this reform, the *Outline of China's National Plan for Medium and Long-term Education Reform and Development (2010-2020)* was published (People's Republic of China, 2010). This policy document outlined key areas of focus for education development over the course of a decade, including education equity, curriculum reforms, teacher training and professional development, higher education, vocational education, and basic education. Emphasis was placed on developing students' thinking to promote creativity and innovation. However, while students' thinking was recognised as an outcome of interest, the document provided little information on what was meant by "thinking" and how it could be achieved, using very vague terms.

In April 2022, more specific directions on cultivating thinking were issued through the curriculum plan and standards for primary and secondary schools by the Ministry of Education of the People's Republic of China. This is the most recent policy aligning with modern demands, including the need for CT skills. The English curriculum standards for secondary schools (MoE, 2022a), for example, stress the development of four core abilities: thinking skills, language abilities, cultural awareness, and learning abilities. Accordingly, classroom teaching methods, assessments, textbooks, and teacher training were revamped to reflect the need to develop these core abilities.

In the English curriculum for secondary schools, thinking skills include interpretation, analysis, evaluation, inference, explanation, deduction, assumption, and self-regulation. These are all elements of CT skills, yet the term "critical thinking" is not explicitly used in the document (MoE, 2022a). Interestingly, the term "critical thinking" is not explicitly used in any government document or policy (Jiang, 2013). This omission may be due to the misleading translation of "critical," which implies criticism or negative thinking (Guo & O'Sullivan, 2012).

In summary, there is a growing policy interest in developing higher-order thinking among students in China. The ambition is for education to grow in line with China's economic and technological advancements. For China to compete in the international arena, it must develop a highly educated population compatible with the demands of the 21st century. This context provides the foundation for my thesis and the focus on CT.

4.2 Critical thinking education in China

While policymakers in China may have the ambition to develop a highly educated population with a focus on CT skills, in practice, the teaching of CT is still immature. CT education in China has largely been confined to higher education. According to Dong (2015), since the late 1990s, CT was introduced in higher education as part of courses on logic. Much of it was ad hoc, with university lecturers translating course materials from the West and calling for more formal teaching of CT (e.g. Chan & Wong, 1999). It was not until 2003 that CT courses in higher education gained traction with the introduction of the Logic and Critical Thinking course at China Youth University for Political Sciences and Peking University. A year later, the Chinese University of Politics and Law and Renmin University of China also launched CT courses. All four prestigious universities are based in Beijing, the capital of China.

New textbooks were written specifically for the CT course. For example, the textbook by Gu and Liu (2006) was highly recommended by the MoE, but most of its teaching materials were borrowed from the West, with little reference to the Chinese context (Dong, 2015). To address this, Dong (2010) published a Chinese textbook that included many practical examples from China. He also taught the CT course to students at Qiming College in Huazhong University of Science and Technology (HUST). Students reported that the course taught them to analyse issues systematically, reflect on their thinking regularly, and provided them with a new perspective on considering issues (HUST, 2011). While this course was considered successful, it was limited to one college where students were highly talented. To expose more students to CT education, Shantou University also implemented a CT course for all freshmen (Dong, 2015).

Some universities, such as Peking University and Tsinghua University, also tried to embed CT into general education modules. The aim was to develop CT and intellectual competencies rather than rote memorisation to pass exams. However, the CT component in these general education courses tended to be theoretical and knowledge-based (Jiang, 2013). For example, the module “Philosophy and Critical Thinking” provided by Fudan College merely introduced students to the thoughts and philosophies of Chinese and Western philosophers (Jiang, 2013). There was a lack of practical

exercises in CT skills. Therefore, despite advances among some universities to incorporate CT teaching into their curriculum, Chinese students are still primarily motivated by passing exams and securing good jobs and high salaries (Jiang, 2013).

In addition to higher education institutions, a small number of schools in China have also attempted to teach CT. The primary school attached to Huazhong University of Science and Technology, an urban primary school, initiated a CT course and trialled its impact (Li, 2017; Zhai, 2015). The promotion of CT was also embedded in specific disciplines, including English language, science, music, and fine arts. There was initial resistance from teachers to teach CT (Zhai, 2015), but after three years, CT became a natural part of the curriculum across all subjects. Interestingly, despite the high-stakes assessment pressure, CT education has been undertaken in Zhonghua High School, an urban high school in eastern China. In this school, CT was embedded in all high school subjects, including Chinese, Maths, English language, Biology, Physics, Chemistry, Politics, History, and Geography (Xu, 2017). These schools are held as exemplary in teaching CT and are all situated in big cities.

Despite the claimed success of these exemplary schools, the teaching of CT in Chinese secondary schools remains *ad hoc*, sporadic, unstructured, and unsystematic. Classroom pedagogies have largely remained unchanged (Dello-Iacovo, 2009). The extent to which CT is emphasised varies between schools and regions. While a few schools have incorporated CT into regular teaching (Zhai, 2015), many schools, especially those in rural areas, still rely heavily on traditional teacher-centred approaches. Classroom teaching remains didactic, with passing exams and rote learning still dominating the Chinese schooling system (Dello-Iacovo, 2009). These challenges impact the implementation of CT education in China (see 4.4 for details).

4.3 Teacher training for critical thinking education

The education reforms in China necessitated changes in teacher preparation to meet the demands of the new curriculum standards. In 2011, over 70 university faculties attended the first national conference for CT education. Following the conference, the Association for Critical and Creative Thinking Education was formed to formalise CT education in China. This association has provided teacher training using Dong's (2010)

textbook, involving activities such as case analysis, group discussion, and Socratic questioning (Zheng, 2017). Additionally, Shantou University launched a training programme focused on inquiry-based critical pedagogy in 2012.

Despite these efforts, the teacher training courses have not been widely popular, facing several challenges. First, only a few teachers signed up for the courses, with some participating due to administrative requirements rather than genuine interest (Zou & Lee, 2023). Secondly, these training courses often attempt to cover a vast amount of content within a limited time frame (Zheng, 2017). There is also a lack of continued support for teachers after the initial training. Thirdly, most of the training focuses on the types of CT, its historical development, and the theoretical and philosophical basis of CT, rather than on practical methods for teaching CT in schools (Zheng, 2017). Consequently, despite the availability of such training courses, teachers are still not adequately equipped to teach CT in the classroom.

Teaching CT involves good questioning techniques, such as asking students for clarifications, explanations, and offering alternative viewpoints. Chinese teachers are not familiar with such techniques, as their traditional role has been to disseminate information and provide answers. There is a distinction between teaching what CT is and developing CT skills. While Chinese teachers may be adept at explaining what CT is, they are not particularly trained to develop those skills in students (Zou & Lee, 2023). Currently, there are no teaching resources specifically developed for teachers to use. Teachers are largely left to their own devices to develop these teaching materials, and naturally, with little to no experience, these materials are often either inappropriate or not genuinely focused on CT.

Overall, while there is growing recognition of the importance of CT in Chinese education, its integration into the curriculum and teacher training is still a work in progress. The implementation of these training programmes can vary, and not all teachers may receive comprehensive training in this area.

4.4 Challenges in implementing critical thinking in schools in China

The Chinese government has ambitious plans for education reforms to upskill its citizens to meet the demands of economic and technological development in the 21st century. Teachers have been encouraged to develop lessons and activities that promote analytical and logical thinking and problem-solving. Textbooks have been revised, and teachers are encouraged to facilitate classroom discussions and debates. However, realising the aims of these reforms has encountered several challenges.

First, the Chinese education system is deeply embedded in thousands of years of Confucian teaching, which values respect for authority and advocacy for conformity. This often conflicts with the Western philosophy of Socrates and the Socratic method of questioning, which many believe to be the early roots of CT. Chinese students are taught from a young age to respect teachers as authority figures and accept everything that teachers tell them without question (Jiang, 2013). Challenging or questioning teachers is considered disrespectful and is strongly discouraged.

Secondly, as explained above, teachers themselves have not been adequately trained to teach CT. Teachers often struggle with the new pedagogy advocated by CT education (Zou & Lee, 2023). Teaching in China is very teacher-centred, whereas the Socratic method of teaching, upon which CT is based, involves asking probing questions to encourage students to question their own beliefs, assumptions, and biases. This is challenging for many Chinese teachers who are more accustomed to providing answers than asking questions.

Thirdly, the education system in China remains very exam-oriented despite the national plans for education reforms for the 21st century. This has proved to be a major obstacle. China's university entrance examination, the Gaokao, is the most important assessment in any student's academic career, determining success in gaining entry to top universities. This exam covers a range of subjects and is very content-based. Teachers and students place a lot of emphasis on learning the content of the subjects examined. As CT is not explicitly examined (even though its skills are useful in learning a number of subjects), it is not emphasised in school. Teachers do not see the relevance of CT in an exam-focused system where rote memorisation is still encouraged. This results in

the problem of teaching to the test. In other words, if the content is not tested, it will not be taught. This explains why CT is not given prominence in classroom teaching despite researchers calling for CT to be taught in schools from an early age (Kuhn, 1999).

Finally, China's education policy context presents a complex backdrop for the implementation of CT as a desired educational outcome. While CT is highlighted in policy documents as an important cognitive goal, it is not explicitly mandated as a requirement in teaching practices for secondary schools. The policy specifies CT as an aspirational outcome but leaves schools and educators considerable discretion in determining how to achieve it. This flexibility enables diverse approaches, such as offering extracurricular CT activities or involving parents in fostering CT skills at home. Infusing CT into the regular curriculum is one option among many. Moreover, supporting actions that could align with CT improvement, particularly through the adaptation of academic exam content, remain limited. In an exam-oriented system where teachers and schools prioritise testable outcomes, this lack of alignment reduces the incentive to integrate CT meaningfully into everyday teaching practices. These factors help explain why current CT teaching practices in schools still remain underdeveloped and inconsistent.

4.5 Chapter summary

This chapter provides the context for delivering CT teaching in China and justifies the necessity of the randomised controlled trial. It presents that while there is a growing emphasis on CT in policy and government documents, the practice is still sluggish. This may be due to traditional values on authority and conformity, teachers lacking support for CT teaching, and a focus on teaching to the test. Recognising the need to navigate cultural values, provide better support for teachers, and shift focus from test-based teaching, this thesis contributes to CT education in China by evaluating the effectiveness of the infusion method.

Section II Research design and methods

There is a common assumption among Western academics that Chinese students are somehow deficient in critical thinking (CT). However, misinterpreting Chinese students' difficulty in demonstrating CT as a lack of ability could result in a waste of resources. Therefore, this study first examines this assumption through a review of studies comparing the CT skills, dispositions, and styles of Chinese students with those of other nationalities. Chapter 5 presents the rationale for conducting a systematic review (SR) and details the processes of searching, screening, extracting, evaluating, and synthesising the relevant literature.

To assess the impact of EnglishFusion on Chinese secondary students' CT skills and academic attainment, a randomised controlled trial (RCT) was conducted. Prior to the main trial, a pilot study was conducted first to assess the feasibility of the research design. Chapter 6 describes the key components of the RCT, including the intervention, case selection and allocation, data collection and measurement. The methods of data analysis and process evaluation are also discussed.

Chapter 5 Systematic review

This chapter presents the rationale for employing a systematic review (SR) to address the research questions, followed by a detailed account of the search strategy, screening process, data extraction, quality assessment, and evidence synthesis.

5.1 Research aim and questions

Previous literature reviews indicate a lack of consensus on the actual CT performance of Chinese students (e.g. Atkinson, 1997; Paton, 2005). The aim of this systematic review is to synthesise existing evidence regarding the CT abilities of Chinese students to evaluate the common assumption about their perceived lack of criticality. To achieve this objective, the research question is framed as follows:

What is the evidence on Chinese students' critical thinking compared with students of other nationalities?

In conjunction with this primary research question, three sub-questions are posed:

1. How do Chinese students' critical thinking skills compare with those of other nationalities?
2. How do Chinese students' critical thinking dispositions compare with those of other nationalities?
3. How do Chinese students' critical thinking styles compare with those of other nationalities?

5.2 Rationale for a systematic review

A systematic review provides evidence-based answers to research questions within a specific field through comprehensive searching, criteria-based selection, critical evaluation, and unbiased analysis (Boland, Cherry, & Dickson, 2017; Klassen, Jadad, & Moher, 1998). This method follows a series of general stages, including identification, screening, and inclusion, thereby explicitly delivering key information and enhancing the transparency of research (Boland et al., 2017; Hammersley, 2020). Additionally, it allows for an in-depth analysis of existing literature, particularly valuable when there are disputes on a topic (Siddaway et al., 2019; Petticrew & Roberts, 2006). Given the

lack of consensus on the CT performance of Chinese students (e.g. Atkinson, 1997; Paton, 2005), a systematic review is an appropriate method for this research.

However, potential challenges in conducting a systematic review must be acknowledged. Balancing the quantity and quality of available literature is important (Zawacki-Richter, Kerres, Bedenlier, Bond, & Buntins, 2020), and pilot testing or a scoping search can be beneficial (Boland et al., 2017). To this end, a scoping search on the CT performance of Chinese students was conducted prior to the formal systematic review.

Ethical issues also pose a potential threat, even though they are often not explicitly discussed in many systematic reviews (Suri, 2020). Although systematic reviews are typically considered low risk in terms of ethical considerations, this does not negate the relevance of ethical issues. Due to the growing popularity of systematic reviews across various disciplines, ethical considerations warrant greater attention (Harlen & Crick, 2004; Suri, 2020). Ethical considerations in systematic reviews should be reflectively engaged throughout the entire review process (Suri, 2020). For example, during the data extraction stage, it is crucial to critically explore how ethical issues might affect research findings in original studies. Ethical information, including conflicts of interest, funding sources, and authors' self-reflection on ethics, should be recorded. Even if limited ethical information is provided in the papers, evaluating the quality of the research design, the claims made, the comprehensiveness of data presentation, and whether the conclusions are warranted by the data are all ethical issues that impact the trustworthiness of the findings. These details must be clearly acknowledged in the reports. Above all, the review itself should be of the highest quality, drawing logical conclusions based on strong evidence (Gorard, 2021).

Overall, a systematic review is advantageous for addressing the specific research questions due to its strengths in transparency and its ability to manage debates effectively.

5.3 Searching strategy

The initial stage of the systematic review involves identifying all potentially relevant literature. In this review, studies are identified through online databases and manual searching. Various methods are employed to retrieve relevant reports.

Online database searching

Given that the research topic falls within the field of social sciences, including education and psychology, relevant studies were sought in social science databases and search engines hosting such databases. The EBSCO host search engine, for instance, grants access to a variety of databases, e-journals, and e-books in education and psychology (e.g. APA PsycInfo), as well as social work. These databases are valuable for identifying journal articles and other publications pertaining to the specific topic within the subject areas covered by each database. ProQuest was also included as it covers Masters' dissertations and PhD theses, ensuring the inclusion of high-quality unpublished work in the review, thus enhancing its comprehensiveness. This distinguishes the current review from previous ones conducted on the same topic. Additional databases used in this structured review include Applied Social Sciences Index & Abstracts (ASSIA), Sage Journals, Scopus, Web of Science, and Wiley online library.

Following the selection of suitable databases, a set of key search terms was developed. These terms directly correspond to the research questions, with a focus on “critical thinking” and “Chinese students.” The keywords used in the search were as follows:

(“critical thinking” OR “think critically” OR “critical reasoning” OR “thinking skill*”)
AND (China OR Chinese)
AND (student* OR learner* OR pupil*)

These keywords are then applied and adjusted as necessary to accommodate the specific search functionalities and idiosyncrasies of each selected database. By employing this tailored search strategy, the review aims to capture a wide range of literature relevant to the research questions.

The search was confined to studies conducted between 2000 and 2021, aligning with the period of educational reform in China that prioritised CT (Chen & Shi, 2017). This timeframe witnessed an increase in research and publications on CT, providing insight into the impact of the reform on Chinese students' CT capacity. Additionally, the search was restricted to studies published or reported in English or Chinese. Importantly, no other restrictions, such as publication types or document types, were imposed to prevent publication bias. Consequently, the review included all relevant published and unpublished materials. The online database search concluded on 14th January 2022, and further details regarding the search strategy and outcomes are provided in Appendix A.

Manual searching

Recognising the potential limitations of solely relying on online databases, hand searching was incorporated to mitigate publication bias and uncover additional relevant literature (Boland et al., 2017; Newman & Gough, 2020). Hand searching, as advocated by Tawfik et al. (2019), helps retrieve papers not captured by database searches. For instance, Google Scholar serves as a valuable source of grey literature (Hagstrom, Kendall, & Cunningham, 2015), offering access to unpublished materials beyond traditional academic publications. In this study, Google and Google Scholar were hand searched to identify grey literature, thus avoiding an exclusive focus on formally published works. Furthermore, to ensure a comprehensive approach, references cited in the studies identified through electronic database searches were also followed up.

5.4 Screening

Having retrieved relevant reports from both database searching and manual searching, the next step was to import them to EndNote (software designed for managing references).

Before screening the imported studies for relevance, a set of inclusion and exclusion criteria was established. Torgerson (2003) proposed three considerations for defining these criteria in a systematic review: the timeframe, research type, and relevance to specific research questions. Guided by these principles and aligned with the research objectives, the following inclusion and exclusion criteria were developed.

The inclusion criteria

Studies were included if they were:

- Concerned with ethnic Chinese students (including students from Hong Kong, Taiwan, and Macau) and students of other nationalities
- About students in schools or higher education
- Related to the assessment of critical thinking
- Empirical (e.g. not opinion pieces, or guidance manual on how to teach critical thinking to Chinese students)
- Published or reported between 2000 and 2021
- Published or reported in English or Chinese

The inclusion of Hong Kong, Taiwan, Macau, and Mainland China, despite educational differences among these regions, enables a broader and more comprehensive examination of assumptions about Chinese students' CT. These regions share a cultural foundation influenced by Confucian philosophy, which emphasises respect for authority. This cultural norm is often interpreted as a lack of critical awareness when individuals refrain from challenging authority or questioning conventional wisdom. Moreover, the inconsistent definitions of "Chinese students" in discussions about their perceived lack of CT make it interesting to adopt broader inclusion criteria. This approach ensures that no important studies that could influence the assumption are overlooked. To address potential ambiguity, the specific groups of Chinese students are clarified in the results chapter (see Chapter 7).

The exclusion criteria

Studies were excluded if they:

- Focused solely on assessing the critical thinking of Chinese students with no comparison with other nationals
- Were not about students in schools or higher education (e.g. there were several studies that examined the critical thinking skills of individuals in different occupations)
- Were theoretical pieces

- Were not primary research
- Did not have measurable outcomes of critical thinking (critical thinking skills, critical thinking dispositions or critical thinking style)
- Based on participants' self-report (i.e. subjective opinions or individual experiences)

For transparency and consistency, the screening process adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (Page et al., 2021). This approach facilitates the systematic recording of the number of research reports identified through database searches, the number included/excluded based on predetermined criteria, the number screened, and the final number retained for synthesis (see Chapter 7). Compared to the 2009 version (Liberati et al., 2009), the updated PRISMA guidelines offer improved transparency in reporting, enabling readers to better assess the credibility of the research findings (Page et al., 2021).

5.5 Data extraction

Following the inclusion of relevant studies, data extraction was conducted to retrieve key information about each study's research design, sampling size, sampling strategy, outcome measures, missing data, method of analysis, and results. This comprehensive summary of information informs the assessment of the strength of evidence, distinguishing this review from previous ones on this topic. Most previous reviews typically do not evaluate the trustworthiness of findings by systematically weighing the research evidence in terms of threats to validity. By incorporating this aspect into the review process, a more robust assessment of the evidence base is achieved. A detailed data extraction table is provided in Appendix B. The table includes the following items:

Study characteristics

- Author(s)
- The year of publication
- Research focus: different aspects of CT

Study design

- Is the study comparative?
- Is the study cross-sectional?

- Is the study longitudinal?

Sample

- Sample size in each group
- Clarification of nationalities of samples
- Sampling strategy
- Setting: discipline, institutions, countries
- Level of education

Measuring instrument

- How are outcomes measured?
- Does the study use independent, standardised, and validated tests?
- Is the instrument developed by researchers themselves?
- Is there any modification in the employment of the instrument (e.g. translated to another language; reduction of measuring contents)?

Research findings and results

- CT performance of each group
- Sub-scale of CT measurement outcome
- Overall results stated by authors (e.g. higher level of CT, lower level, mixed or no difference)

Limitations acknowledged by the author(s)

- Issues about generalizability (e.g. scale, sampling strategy, attrition)
- Quality of the use of the instrument
- The demographic information that may influence CT performance (e.g. admission criteria, CT courses)

Rating based on “sieve” (Gorard, 2021, p.94)

- Research design
- Scale
- Missing data
- Measurement quality
- Rating outcomes (from 0* to 4*)

5.6 Quality assessment

Quality assessment is an essential aspect of the systematic review process, as it ensures that the findings are based on robust evidence. The intention of quality assessment is

not to exclude reports with poor quality, but to critically examine their influence on the research findings (See, 2018). Failure to assess the quality of each study and indiscriminately including low-quality evidence in the synthesis can lead to incorrect conclusions (Ahn & Kang, 2018). For this reason, the review does not exclude studies based on research design to ensure that all kinds of evidence are considered. The quality of the research and the appropriateness of the research design in addressing the research questions determine the strength of evidence for each piece of work.

Various tools are available for quality assessment in systematic reviews. For instance, checklists proposed by the Critical Appraisal Skills Programme (2021) are valuable for evaluating healthcare evidence. However, they may be less relevant in the education field. Alternatively, the Cochrane risk-of-bias tool for randomised trials assesses bias arising from various aspects of study conduct and reporting (Higgins et al., 2021). Nonetheless, this tool may be time-consuming to apply (Boland et al., 2017), and consistency across multiple appraisal items can be challenging.

Most critical appraisal tools for systematic reviews are simply checklists focusing on the quality of reporting. For example, the Joanna Briggs Institute (JBI) appraisal of systematic reviews includes questions such as whether the review question was explicitly stated, whether the inclusion criteria were appropriate, and whether the search strategy was suitable (JBI, 2017). However, these checklist criteria may be vague and open to interpretation, particularly when multiple reviewers are involved. Moreover, JBI has formulated appraisal protocols for evaluating the quality of studies across various research designs such as qualitative studies, cohort studies, and randomised controlled trials. While this approach is beneficial, it falls short in cases where a review incorporates studies employing different designs. For instance, studies featuring a limited sample size (e.g. two individuals) and no comparison groups, and relying on self-reported data or perceptions may receive a favourable quality rating if they satisfy all assessment criteria, such as alignment between philosophical perspective and methodology, or adequate representation of participants' voices. Conversely, another study within the same review using large administrative datasets, controlling for confounders and missing data, and incorporating participant perspectives but failing to

state the researcher's philosophical stance may be deemed low in quality. The use of a mix of appraisal tools in such reviews can potentially lead to misleading conclusions.

This study employs a quality appraisal tool called the "sieve", developed by Gorard (2021, p.94), to assess the trustworthiness of findings in each included study. The "sieve" considers various aspects of research designs, primarily focusing on their appropriateness for addressing the research question and addressing potential threats to validity such as attrition or missing data, as well as the quality of outcome measurement. Notably, the authors' reputation and the publication outlets are disregarded, with each piece being evaluated solely based on these specified criteria as outlined in Table 5.1. To ensure inter-rater reliability, each study was rated by two reviewers. In instances of disagreement, consensus was reached through discussion and careful review of the criteria.

Studies are rated according to the "sieve" table. Evaluation begins from the top left, assessing the strength of the research design relative to the research question, and progresses across to evaluate sample size, missing data, and measurement quality sequentially. Ratings either remain constant or decrease when moving from left to right in the table. For instance, if a study employs a robust design for the research question but has a small sample size, the score decreases from 4* to 2*. Even if there is no dropout and a standardised instrument is used to measure outcomes, the overall score still remains 2*. This is because the "sieve" assigns different weight to various aspects of quality assessment. The appropriateness of the research design receives the highest priority, while measurement quality is considered least. If a study fails to consider research design, it receives a rating of 0*, irrespective of sample size or completeness of data. Notably, the "sieve" table does not provide specific numerical thresholds to distinguish between large, medium, and small scales, nor does it offer concrete thresholds for different levels of missing data. They are left vague intentionally, so reviewers are required to exercise judgement based on the specifics of each study (Gorard et al., 2017). For example, a study with 100 randomly assigned participants is considered larger in scale than one with only 50 individuals but smaller than one involving 500 participants. Similarly, an attrition rate under 5% is deemed less likely to impact final outcomes than a rate of 20%.

Table 5.1 A “sieve” to assist with quality assessment

Design	Scale	Missing data	Measurement quality	Rating
Strong design for research question	Large number of cases (per comparison group)	Minimal missing data, no impact on findings	Standardised, independent, reasonably accurate	4*
Good design for research question	Medium number of cases (per comparison group)	Some missing data, possible impact on findings	Standardised, independent, some errors	3*
Weak design for research question	Small number of cases (per comparison group)	Moderate missing data, likely impact on findings	Not standardised or independent, major possible errors	2*
Very weak design for research question	Very small number of cases (per group)	High level of missing data, clear impact on findings	Weak measures, high level of error, or many outcomes	1*
No consideration of design	A trivial scale of study	Hugh amount of missing data, or not reported	Very weak measures	0*

Source: Gorard (2021, p.94)

5.7 Synthesis

The synthesis of included studies is structured around the three dimensions of CT: skills, dispositions and styles. Within each dimension, studies are categorised based on whether they report higher, lower, or similar levels of CT. The strength of evidence for each level is determined by considering both the number of studies and their quality ratings. For instance, if a majority of studies rated 3* yield mixed results, it suggests that the evidence for that dimension is mixed. The study with the highest rating carries the most weight in informing the evidence. In cases where none of the studies receives

a rating above 2* and the distribution of studies across rating levels is even, it indicates unclear evidence. Likewise, if the majority of studies receive a rating of 1*, and all of them indicate that Chinese students exhibit lower CT skills, we cannot confidently conclude that Chinese students indeed possess lower CT skills due to the weak evidence (as indicated by the quality ratings). Thus, the evidence remains tentative.

5.8 Chapter summary

In summary, this chapter highlights the suitability of employing a systematic review methodology to establish evidence for the claims that Chinese students are lacking in CT. It explains the searching strategies, criteria for inclusion or exclusion of studies, types of data abstracted from included studies, the quality assessment tool employed, and the methods employed for synthesis.

Chapter 6 Primary research

This chapter discusses the empirical study's methodology, detailing key aspects of the randomised controlled trial (RCT), such as the sample, randomisation, and intervention. Prior to the main trial, a pilot study was conducted to test the feasibility of the intervention materials and to collect formative feedback from teachers and students, which would inform improvements for the main trial. Revisions to the intervention and questionnaire instruments updated for the main trial are presented. This chapter also discusses the outcome measures and the methods of analysis. To understand why and how the intervention works or does not, a process evaluation was also conducted, the findings of which will supplement those of the impact evaluation. Finally, ethical issues related to the primary research are addressed, and pragmatic ways of dealing with them are reported.

6.1 Research aims and questions

The main aim of the primary research was to evaluate the impact of infusing critical thinking (CT) into the English curriculum on Chinese students' CT skills and academic attainment. A second aim was to explore whether CT skills can be taught to Chinese secondary students who are not traditionally exposed to CT education. The study will also explore whether training teachers to deliver CT in their lessons has changed their critical awareness and views on CT teaching.

In accordance with these aims, the following research questions are proposed:

1. Can critical thinking skills be taught to Chinese secondary students who are not traditionally exposed to critical thinking?
- 2a. Does EnglishFusion improve Chinese secondary students' critical thinking skills?
- 2b. Does EnglishFusion have a differential impact on the critical thinking skills of sub-groups of students (by age, birth sex, ethnicity, prior academic attainment, prior critical thinking skills, schools, parental involvement in children's education, and home background)?
- 3a. Does EnglishFusion improve Chinese secondary students' academic performance?
- 3b. Does EnglishFusion have a differential impact on the academic attainment of sub-groups of students (by age, birth sex, ethnicity, prior academic attainment, prior

- critical thinking skills, schools, parental involvement in children’s education, and home background)?
4. Does training and teaching EnglishFusion alter teachers’ critical awareness and attitudes towards teaching critical thinking?

The primary aim of the research is to evaluate the impact of infusing CT into the English curriculum (i.e. EnglishFusion) and the development of CT skills. This is a causal question, making an experimental or causal design the most appropriate. Therefore, an RCT is adopted to address the research questions.

6.2 The pilot study

Prior to the main study, a pilot study was conducted. The pilot study serves several purposes — testing the teaching and learning resources, assessing the feasibility of conducting the study in Chinese schools, and evaluating the teachers’ ability to deliver the lessons during regular English classes (Feeley et al., 2009). Additionally, the pilot study aimed to trial the tests and surveys to determine if their format, layout, and level of difficulty were appropriate for the age group of students, and to evaluate the time required to complete these tests and surveys. The pilot helped to fine-tune the training for the main trial and identify potential challenges.

The pilot study was conducted from 2nd November to 29th December 2022 in a public secondary school in Sichuan province, China, selected for convenience. This was a low-resource school located on the outskirts of the capital city, similar to the schools in the main study. It involved two English language teachers, each responsible for teaching one Grade eight class (n = 122 students). One class/teacher was randomly assigned to the experimental condition (n = 61 students) and the other to business-as-usual control (n = 61 students). Despite the challenges posed by COVID-19 during the pilot study period, there were no dropouts.

6.3 The intervention

Brief name

EnglishFusion: Think and Learn

Development of the intervention

EnglishFusion: Think and Learn is an intervention that infused CT in the regular English curriculum where CT is taught explicitly but within the context of the existing curriculum. As there were no readymade lessons for teaching infusion of critical thinking in Chinese contexts, I developed a whole teaching module specially for this study using Elder and Paul's (2020) framework for infusing CT. To prepare for the development of the intervention, I received a teacher training course, "How to Infuse Critical Thinking into Your Instruction" in Spring 2022, provided by the Foundation for Critical Thinking. I also reviewed Nilson's (2021) book "Infusing Critical Thinking into Your Course: A Concrete Practical Approach," which provided the foundational knowledge of CT teaching methods (Fan, 2022).

Some modifications were made to Elder and Paul's (2020) framework to ensure that the intervention was appropriate for Chinese secondary students. First, only those intellectual standards relevant to reading and comprehending texts were included. These included six intellectual standards (i.e. clarity, accuracy, relevance, depth, breadth, logic) and three elements of reasoning (i.e. information, assumptions, inferences). The aim of these lessons was to stimulate students to think critically about daily life issues familiar to them. Second, although the development of CT dispositions is a focus in Elder and Paul's (2020) framework, it was not included here because developing CT dispositions requires at least a year or more, which would not be practical in a three-year PhD programme. Also, attributes of dispositions are often self-reported, and there is no objective measure or test of dispositions. Thirdly, the language in the lessons was simplified taking account of the fact that the students are English language learners.

It was originally planned for six EnglishFusion lessons. Lessons 1 to 4 focused on intellectual standards, specifically clarity and accuracy, relevance, depth and breadth, and logic, while Lessons 5 and 6 covered examining information and identifying assumptions. The pilot study found that the six lessons did not adequately develop student's inference skills. An additional lesson explicitly teaching students to make reasonable inferences was added to the main trial. Six CT tasks using textbook contents

from the students' regular English curriculum were included in the main trial as there were concerns by the teacher that the infusion lessons had no links to the regular English curriculum to provide a stronger integration of CT into the regular English lessons. Pilot students also expressed the need for more EnglishFusion lessons to practise CT skills. Another revision made to the main study involved replacing the homework assignments with in-class activities, as it was observed that students rarely completed the homework due to their heavy schoolwork.

Procedures, activities and processes used in the intervention

EnglishFusion was conducted in two stages (see Table 6.1). The first stage consisted of seven lessons aimed at introducing CT to students. Each lesson included five essential sections: the lead-in activity, CT objectives, presentation, practice, and summary. The lead-in activity was designed to introduce the lesson topic and stimulate students' thinking. This was followed by the lesson objectives so that students knew what they were expected to achieve by the end of the lesson. Key concepts were taught in the presentation section with concrete examples and clear explanations. Students then completed practice exercises and were invited to share their answers or thoughts. Finally, students summarised what they had learned, focusing on understanding how to use thinking skills rather than merely memorising knowledge.

The second stage integrated six CT tasks with the content from the regular English textbook. Based on teachers' English teaching pace, corresponding textbook content was extracted and combined with CT learning.

Table 6.1 Summary of EnglishFusion content at two stages

Stages	EnglishFusion topics
<p style="text-align: center;">Stage one <i>Introduction of CT</i></p>	Lesson 1 Clarity and accuracy
	Lesson 2 Relevance — the straw man fallacy
	Lesson 3 Depth and breadth
	Lesson 4 Logic — correlation and causation
	Lesson 5 Examine information
	Lesson 6 Identify assumptions
	Lesson 7 Make inferences

Stage two <i>Regular CT practice in the English curriculum</i>	Task 1 The Monkey King
	Task 2 Gretel and Hansel
	Task 3 Cultural heritage protection
	Task 4 Pandas
	Task 5 The forum of book for children
	Task 6 Country music

Stage one: introduction of critical thinking

Lesson 1 Clarity and accuracy

This lesson discussed two intellectual standards: clarity and accuracy. Clarity requires clear information, while accuracy ensures that thinking is free from mistakes. Both concepts were introduced with everyday examples. For instance, students were asked to judge whether the instruction “Take these tablets three times a day” was clear and to consider the possible consequences if people misinterpreted it.

Additionally, teachers explained the difference between facts and opinions. Students identified facts and opinions among a set of sentences such as “No two people have the same fingerprints” and “My nose is too long,” and were required to justify their answers. They also wrote a few sentences about facts concerning cats and opinions on holidays.

The final activity involved reading a short diary and evaluating its clarity and accuracy. This practice was also designed to exercise their writing skills, as students were asked to suggest revisions and write improved versions of the diary entry.

Lesson 2 Relevance — the straw man fallacy

Relevance involves relating information to the matter at hand. A common fallacy regarding relevance is the straw man fallacy, where an argument appears to refute a statement but does not address the key point. Students were presented with everyday conversations to identify irrelevant points in people’s arguments and suggest ways to counter straw man arguments. For example, they examined the following conversation:

Headteacher: “The school lunch budget must be examined to cut out waste.”

Parent: “This guy wants to starve our children.”

After identifying straw man fallacies in different contexts, teachers introduced three strategies to combat them and maintain focus on the key point: asking for reasons behind opposing views, providing more details to validate the original statement, and repeating the key point while ignoring the straw man argument. Students were encouraged to think of other examples of straw man fallacies from their lives and share them in class.

Lesson 3 Depth and breadth

Depth in thinking involves revealing complex situations or problems through multiple approaches, such as tracing origins, analysing influencing factors, and discussing possible consequences. Breadth involves considering different stances or facts about an issue. To teach students to think deeply and broadly, two topics, rainy days and shorter school days, were debated in class. The class was divided into two groups: one discussed the advantages of shorter school days, while the other considered the disadvantages. A group competition was incorporated to engage students and increase their enthusiasm, with points awarded for answers and reasonable explanations. The group with the highest score won the competition and received a reward.

After the debate, students learned to make counterarguments using the following model to integrate their thoughts: *“We should ... (argument) because ... (the reason/more information to justify). I know that ... (the opposing argument), but ... (argue against the opposing argument).”* Students then applied this skill to two everyday scenarios, trying to persuade the opposing side.

Lesson 4 Logic — correlation and causation

Correlation and causation are often confused in logical thinking (Rohrer, 2018). In the lead-in activity, teachers presented a fact: “More people die if they sleep in a hospital bed than in their own bed,” and asked students what should be done to reduce deaths. Students discussed and shared their answers without being given the correct answer, as they would revisit this question later.

Specific examples were provided to explain the concepts of correlation and causation. In cases of correlation, there may be a third factor or other possible explanations, and the direction of the relationship is unclear. For instance, students were asked to consider other factors that could influence height. In contrast, a causal relationship has a clear direction and time sequence. In other words, one event must occur first, followed by the outcome. This was illustrated using billiards: if someone hits the white ball, it moves. While the first instance might be a coincidence, repeated occurrences confirm the causation.

Correlation and causation are not the only ways to explain relationships. Three other common explanations were also presented: the cause being the other way around, there being a hidden cause, and simple coincidence. After learning these different explanations, students were asked to analyse and explain the hospital bed example, identifying the illness as the real cause. To link the lesson content to daily life, students engaged in a discussion, sharing their own examples of confusion between correlation and causation and suggesting more reasonable explanations.

Lesson 5 Examine information

In today's information-saturated world, much of the information available is fake or false. The ability to critically examine information is essential for combating disinformation (Horn & Veermans, 2019). This lesson introduced three key aspects for evaluating information: checking the source, evaluating the content, and comparing it with other information. Unlike previous lessons that introduced various intellectual standards, this lesson focused on applying these criteria in practice.

In the pilot study, this lesson included two passages about research on eating eggs giving conflicting results about the risk of heart disease. However, students found it difficult to understand and the pilot teacher suggested replacing these long and abstract materials with a new and more relevant example. The teacher disliked the lead-in section of the lesson on examining information because it was too lengthy, and some information seemed repetitive and unrelated to students' daily lives. Thus, one passage was removed and the number of questions introducing the topic was reduced in the

main trial. The egg example was retained because it served as a good exercise for students to evaluate information.

Additionally, the pilot Lesson 5 instructed students to evaluate information from five aspects: currency, relevance, authority, accuracy, and purpose. While the pilot teacher emphasised that the authority standard did not imply that experts were always right, this criterion could easily be misleading. As the previous four lessons had introduced some intellectual standards based on the Elder and Paul (2020) framework, the fifth lesson in the main trial was updated to incorporate these earlier standards. This allowed teachers to guide students in reviewing what they had learned and modelling how to assess information based on these aspects.

After modifications informed by the pilot study, this lesson was taught in the main trial as follows. To start the lesson, a YouTube video was shown, depicting an eagle snatching a child in a park. After watching the video, students were required to summarise its content and judge whether they trusted it, explaining their reasons. Teachers guided students to consider the author, the platform, and content anomalies such as the disappearance of a wing and a strange shadow. It was then revealed that the video was fake, created for an animation project. This was to show students the prevalence of fake information on social media.

Next, students learned how to check the information source by examining a news article about a mysterious seven-foot creature spotted in Argentina. They were instructed to evaluate the author, other stories on the website, the layout, and the image of the news. They concluded that the story was likely fake. Additionally, students read a research report claiming that people who eat one egg per day are more likely to develop heart disease. They were taught to use intellectual standards, including clarity, accuracy, and logic, to judge the trustworthiness of the report's findings. Another piece of information about age and heart disease was provided for students to compare and determine whether to trust the research findings.

Lesson 6 Identify assumptions

An assumption is an idea that people often take for granted. It acts as a bridge between information and conclusion. Identifying assumptions is a crucial part of argument analysis and evaluation (Elder & Paul, 2020; Yanchar & Slife, 2004). The lesson began with a lead-in activity focused on gender stereotypes, making students realise that people often make implicit assumptions about genders. For instance, students were asked to decide whether statements like “likes pink,” “has short hair,” and “has a driving licence” applied to an eight-year-old girl, a nine-year-old boy, both, or neither.

Following this, the concept of assumptions was explained, and students participated in a group discussion exercise to identify implicit assumptions in various statements. However, merely identifying assumptions is insufficient; it is also necessary to evaluate their validity. For instance, students were instructed to identify the assumption in the statement “Old Tom is a whale, so Old Tom is a mammal” and then evaluate whether the assumption was true.

A practice activity identifying and evaluating assumptions was given (see Table 6.2). Teachers modelled how to complete the first row, explaining that while the conclusion might not be certain, the assumption should align with it. Students then discussed the remaining rows and shared their answers in class. To help students apply this skill in daily contexts, they were tasked with selecting two pieces of information from their lives, making conclusions, identifying the assumptions linking the information and conclusion, and evaluating their validity.

Table 6.2 Practice with conclusions and assumptions

Information	Conclusion	Assumption	Is the assumption true? Why?
Your friend does exercise a lot.			
Your best friend has not chatted with you for several days.			
You see one of your classmates taking medicine.			

Your teacher asks you to come to the office.			
--	--	--	--

Lesson 7 Make inferences

Making inferences is a reasoning element that explores implicit meanings or determines what follows next (Elder & Paul, 2020). This lesson was not taught in the pilot study but was newly developed for the main trial. It consisted of two parts: making reasonable inferences from texts and introducing three types of inference rules.

Students read a short passage about a common scenario in life, attempting to infer the location of the event and identify supporting clues to justify their answers. Although the exact place was not mentioned, students could infer that it was a restaurant based on keywords such as “menu” and “table.” This type of inference is sometimes assessed in students’ regular English reading tests.

Three kinds of inference rules including modus ponens, hypothetical syllogism, and modus tollens, were taught using Venn diagrams. Students were instructed to assume the information from the questions was correct and to choose the correct conclusions. For example, after reading the information, “All dogs are animals. Bobby is a dog,” students were given three options:

- A. Bobby might be an animal.*
- B. Bobby is an animal.*
- C. Bobby cannot be an animal.*

Teachers used Venn diagrams to illustrate the relationships and explained modus ponens. Variants of these inference rules were also discussed in class.

Since this was the last lesson of the first stage, students were invited to share their ideas about the differences between CT lessons and usual English lessons. They were also informed that they would encounter concepts learned from these seven CT lessons in their future English studies.

Stage two: regular critical thinking practice in the English curriculum

Task 1 The Monkey King

Students were divided into two groups for the first CT task. One group argued for the positive characteristics of the Monkey King, while the other argued for the negative characteristics based on their background knowledge. All their answers needed to be supported by examples or explanations rather than personal feelings. After the discussion, students read a textbook passage about the Monkey King and examined whether the author had adopted a balanced view. Additionally, students revisited the concept of correlation and causation using the Monkey King example.

Task 2 Gretel and Hansel

This section of the textbook introduced the story of Gretel and Hansel. Three scenarios were selected and paraphrased to improve inference skills. Each question provided three options to facilitate students' responses. Teachers were encouraged to use Venn diagrams to clearly explain the inference rules to their students.

Task 3 Cultural heritage protection

This CT task was based on the inference testing questions from the Watson-Glaser test. The reading material for this task was revised to focus on the protection of cultural heritage, as the textbook had already introduced students to several places of interest such as the Great Wall and the Palace Museum. Unlike the previous task that used inference rules to solve problems, this task aimed to teach students to evaluate the strength of the link between existing evidence and proposed conclusions. For example, students learned that more information should be required if the current evidence provided no basis for the conclusion.

Task 4 Pandas

Students were required to compare and examine two pieces of information about pandas. One source was from the National Geographic Kids magazine website, and the other was an article about pandas from their textbook. Students were guided to evaluate the information about pandas' characteristics by answering questions regarding the information source, purpose, and other credible evidence needed.

Task 5 The forum of book for children

This task was adapted from the reading literacy items in Programme for International Student Assessment (PISA, 2018). A forum recommending books for children was presented to the students, along with posts by different netizens. Students read and discussed all the posts from the forum. Specifically, they interpreted literal meanings, determined the relevance of each post, inferred the purpose of some responses, and assessed the quality and credibility of the recommendations. They also considered open questions such as whether it was necessary to post queries on the forum and what alternative solutions they could suggest.

Task 6 Country music

After learning about country music in their usual English class, students were more familiar with its characteristics. In the final CT task, they read a short passage comparing country music and hip-hop and tried to identify the correct assumption made from the material. Three options were provided for selection, and students were required to justify their answers.

Teaching and learning resources

The teaching and learning resources developed for the intervention included student handouts with lesson activities, lesson slides, and lesson plans. These were provided to support teachers in CT instruction. An example of the teaching and learning materials for Lesson 5 can be found in Appendix C.

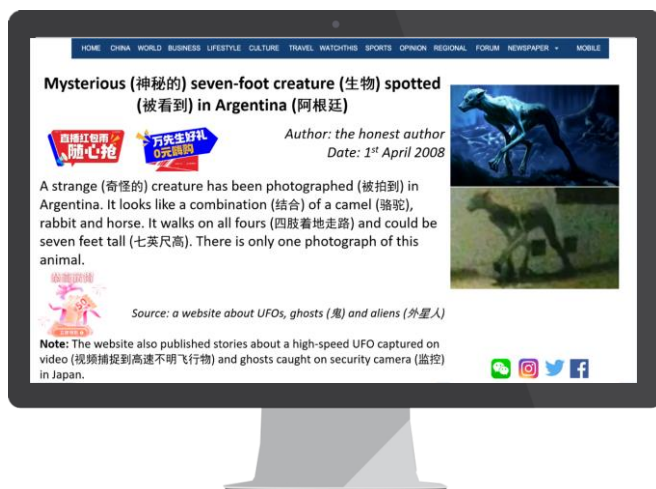
Initially, in the pilot study, there were no **student handouts**. The pilot teacher suggested that it would be helpful to provide printed student handouts to accompany the slides during EnglishFusion class. These handouts supplemented materials from their textbook. It also helped reduce over-reliance on slides, an issue highlighted by the pilot teacher. With the student handouts, students could read the questions and materials and take notes during discussions. Therefore, student handouts were developed for the main trial to facilitate better engagement in student activities and stimulate discussions.

All lessons, except Lesson 4, were accompanied by printed student handouts. These handouts included examples for teachers to model and explain, as well as exercises for students to complete individually or in groups. Blank spaces were provided for students to read, discuss, and write during in-class activities. Pictures were added to match the lesson content and to make it more attractive. These handouts were in black and white to reduce cost. Only essential handouts were printed, meaning those for Lesson 4 and all tasks in Stage 2 were not provided, as students could read the content and complete exercises using the slides. Teachers received student handouts one week in advance to prepare for their CT lessons, while students received each lesson handout only at the beginning of the scheduled lesson. This approach aimed to attract students' interest and curiosity about the lesson content. Figure 6.1 is an extract of the student handout for Lesson 5.

Figure 6.1 An extract of the student handout for Lesson 5

Lesson 5 Examine (审查) the information

Information 1. Check the source (来源): the website story



1. Who reported (报道) the story? Have you heard of them before?

2. What other stories did they share? Do these stories seem believable (可信的)? Why?

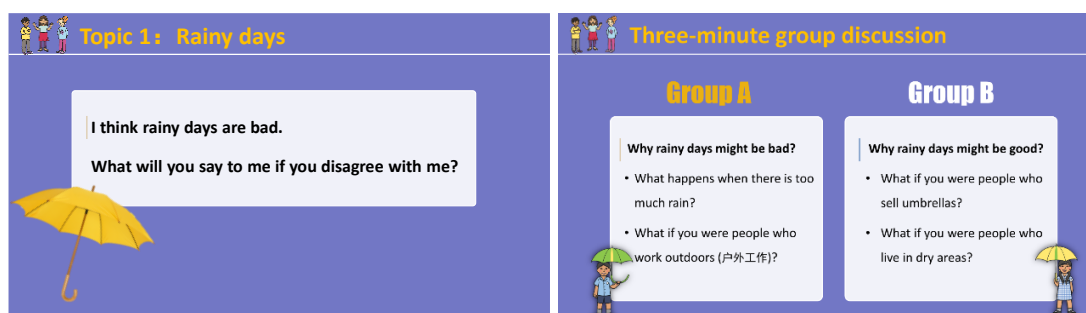
3. Does the website look normal (正常的)? Why?



Given that using slides is common in regular classes and all classrooms were equipped with computers and screens, **lesson and task slides** were provided for teachers to directly use in class. It could also reduce teachers' preparation workload. The lesson slides were improved based on feedback from the pilot study. The pilot teacher suggested that the slides should specify how and for how long each activity should be conducted. Although this information was included in the teaching plan, the teacher thought it would be more effective if the slides also contained these details. It was hoped that the adjustment would help teachers manage EnglishFusion lessons more efficiently and ensure the successful completion of activities.

Moreover, as suggested by both students and the pilot teacher, adding more pictures to the slides would make them more attractive. Therefore, in the main trial, the slide content was improved, using authentic website pages to help students connect the lesson content with their life experiences. Pictures from the student handouts were incorporated into the slides to illustrate abstract concepts and examples vividly. These pictures were in Portable Network Graphics (PNG) format to ensure clarity and avoid distractions from irrelevant information or unclear images. Online videos (Lessons 5 & 7) were also embedded in the slides to aid class discussions. Various transitions and animations were used to maintain student engagement in CT learning. An example of lesson slides for one activity in Lesson 3 is in Figure 6.2.

Figure 6.2 An example of lesson slides for an activity in Lesson 3



Accompanied by handouts and slides, **lesson plans** were sent to the teachers. These plans helped the experimental teachers understand the content to be taught, the methods of delivery, and how the slides would be used during class. The lesson plans included suggested procedures and the estimated duration of each section. Each slide was


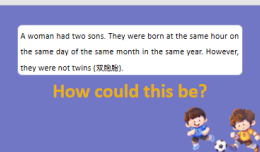


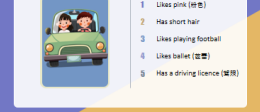
incorporated next to the corresponding teaching content. The lesson plans detailed teachers' instructions, including explanations and questions designed to prompt students' thinking. While there were no standard answers for some open-ended questions, indicative answers were provided to show teachers what was expected from students during CT instruction. It was clearly stated in the lesson plans that these answers were not intended to restrict students' thinking but to encourage them to explore different perspectives. If students struggled to answer some questions, teachers could use the provided answers to model and explain. Additionally, several notes about CT teaching were consistently mentioned in the lesson plans to help teachers become increasingly familiar with CT instruction. For example, teachers were reminded to always ask students to provide reasons or elaborate on their answers. An extract of the lesson plans is shown in Figure 6.3.

Among all these materials, key words and phrases were translated into Chinese to accommodate students' English language skills. In the pilot study, there were not many Chinese translations in lesson slides, so the teacher had to translate all new words and sentences before students could complete tasks. According to the class observations, different levels of English proficiency among students could affect their understanding and engagement. As most students in the primary research were beginners in learning English, more Chinese translations and fewer technical terms were included in all intervention materials to ensure student comprehension. Teachers were also encouraged to use additional Chinese translations in the slides if they anticipated that students might struggle with understanding the content.

These teaching and learning materials were originally sent to the pilot teacher all at once. However, this method was found to be daunting and added pressure on an already overworked teacher. Therefore, in the main study, they were not packaged together at the start of the intervention. Each lesson's teaching and learning materials were sent one week in advance to experimental teachers.

Figure 6.3 An extract of the lesson plan for Lesson 6

Lesson 6: Identify assumptions (识别假设)

<p>Lead-in (6 minutes)</p> <p>1</p> 	<p>1. Two sons</p> <p>Teacher: A woman had two sons. They were born at the same hour on the same day of the same month in the same year. However, they were not twins. How could this be?</p>
<p>2</p> 	<p>(Students give as many answers as possible.)</p> <p>Possible answers: They're triplets (三胞胎)—there is a third child; they were adopted (被收养); ...</p>
<p>3</p> 	<p>Teacher: This example simply states there are two sons born at the same time and did not in any way indicate that they are twins. Yet, in an attempt to quickly get an answer, most people are likely to assume (假定) that two children are twins. That assumption is unwarranted (不合理的).</p>
<p>3</p> <p>★</p> 	<p>2. Emily and Jack</p> <p>Teacher: Emily is an 8-year-old girl and Jack is a 9-year-old boy. Identify which of the following statements are about Emily, which ones are about Jack, which could be about both, and which could be about neither. Let's look at the first statement. Is it about Emily, Jack, both or neither? (Go through each statement)</p>
<p>4</p> <p>★</p> 	<p>1. likes pink (粉色) 2. has short hair 3. likes playing football 4. likes ballet (芭蕾) 5. have a driving licence (驾照)</p>
	<p>(Students answer)</p> <p>Teacher: Can you explain your answer? Why do you think so?</p> <p>(Students give reasons)</p> <p>Teacher: Any other thoughts? (Ask for different ideas and reasons.)</p> <p>(Students answer)</p>

Duration and dosage

EnglishFusion was delivered by English language teachers in the participating schools as part of their regular English curriculum in the students' usual English language classrooms. The intervention was conducted from March to May 2023 in two stages. In the first stage, CT lessons were taught once a week, with each lesson lasting 40-45 minutes. After seven continuous weeks of CT lessons, there was a two-week break for the mid-term test, collectively scheduled by the local Department of Education. Following the test, students had a holiday for International Workers' Day.

The second stage of the intervention lasted about one month. Since CT tasks were linked to the regular English teaching content and were expected to take around 10-20 minutes (less than a lesson period) to complete each task, teachers were asked to integrate them based on their individual teaching pace. This meant that teachers determined when to complete the six CT tasks.

Training of teachers

Prior to the delivery of the lessons, teachers were trained to use EnglishFusion to ensure that they implemented the lessons and tasks as intended. In the pilot study, the training was conducted online due to the COVID-19 pandemic. A researcher-developed workbook was sent to the pilot teacher to introduce the background of CT teaching in China. The six CT lessons and the course logic were clarified, and teaching suggestions were provided to facilitate lesson preparation. Notably, the workbook was written in English, which could have required the teacher to spend additional time reading and comprehending it, as English is not her first language. Therefore, for the main trial, the workbook was modified to a Chinese version and focused more on practical use.

The pilot teacher training was conducted informally. The experimental teacher suggested she could self-train by reading the workbook and the provided teaching resources. It was agreed that she would contact me if any problems arose. This approach was understandable given her heavy workload and tight schedule. During the intervention period, there was ongoing communication regarding feedback on the implementation of the CT lessons, including student reactions, teaching materials, and teaching styles. Nonetheless, the lack of modelling and rehearsal before class

contributed to the low fidelity of the intervention. This indicates that the self-learning and informal online communication mode of teacher training did not adequately support teachers' delivery of infusion CT lessons. To better prepare teachers for EnglishFusion lessons in the main trial, I conducted one formal teacher training session, supplemented by weekly informal follow-up sessions throughout the intervention period, all of which were conducted face-to-face.

The formal teacher training session was conducted on 2nd March 2023. All 11 experimental teachers from the four schools participated in this face-to-face training, which lasted for 2.5 hours and took place at School A. This training ensured that teachers understood their roles and responsibilities in CT teaching and were prepared to follow the planned EnglishFusion.

I delivered the formal training, covering the following aspects:

- Clarification of CT: A brief explanation of CT, the project background, and its purpose.
- Explanation of teachers' tasks: Detailed information about the experimental teachers' responsibilities, including teaching seven lessons and six tasks according to a scheduled timetable and maintaining the confidentiality of intervention-related materials.
- Teaching strategies: Guidance on using the teaching materials, classroom strategies (e.g. asking for explanations of students' answers, avoiding directly giving answers), and questions to inspire further thinking (e.g. "Can you give us an example?", "Why did you say that?", "How could we check whether that was true?").
- Lesson demonstration: A demonstration of the second lesson, where teachers acted as students and discussed questions.
- Teaching practice: Practice sessions for the first and third lessons with other experimental teachers.
- Instructional video: A video of a dialogue class to facilitate understanding of instructional methods.

- Question and answer session: An opportunity for teachers to ask questions and receive clarifications.

A teacher training slide covering these key procedures was used during the formal training session. Additionally, to support the implementation of the intervention, teachers received a researcher-developed workbook, consistent with the training content and written in Chinese to facilitate understanding. Teachers were allowed to keep this workbook.

Alongside the workbook, teachers were provided with teaching materials for Lessons 1-3, including slides, lesson plans, and student handouts. To explain the use of these materials, I conducted a demonstration of Lesson 2, where teachers, acting as students, read the student handout and answered activity questions. This allowed them to experience the CT content from a student's perspective. Following this, teachers were encouraged to form groups and practice Lessons 1 and 3.

An additional training resource included a video recording of a dialogue class in China. Unlike traditional classes, this dialogue lesson encouraged independent thinking and provided many opportunities for group discussions and idea-sharing. This video illustrated how to create an open classroom atmosphere and the expected role of teachers in CT lessons.

In addition to the formal training session, I conducted **weekly informal follow-up training** sessions during the intervention period. I sent teaching materials for each lesson and explained the content to the experimental teachers one week prior to the scheduled teaching date, ensuring they had sufficient time to prepare for their CT teaching. Teachers from the same school often formed groups to prepare together and share ideas. They discussed the content to ensure they understood it and were familiar with the teaching process. During the teaching week, face-to-face meetings were held to address any problems or concerns regarding the CT lesson. Attention to specific matters was stressed based on observations from other experimental classes. After each CT lesson, feedback and reflections were collected from both teachers and students. Teachers reflected on class interactions and sought improvements for future CT lessons,

while I provided suggestions on how to encourage more student participation and ask prompt questions.

6.4 Trial design

The research design employed to address the research questions in the primary study is a two-armed cluster randomised controlled trial (RCT). Given that the research questions are causal in nature, an experimental design capable of establishing causation is deemed the most appropriate (Gorard, 2013). RCTs are often regarded as the gold standard because they can control for systematic differences between the groups being compared. Although no study on its own can prove causality, randomisation reduces bias and provides a robust method for examining the cause-and-effect relationships between the intervention and the outcome (Hariton & Locascio, 2018). By randomly allocating participants to either experimental or control conditions, randomisation eliminates selection bias and ensures that both known and unknown confounding factors are balanced, making the control group as similar as possible to the treatment group (Gorard, 2013; Torgerson & Torgerson, 2001). The control group, also known as the counterfactual, enables observation of what would have occurred in the absence of the intervention. Therefore, without a control group, it is impossible to determine whether the intervention is effective.

However, this does not imply that the findings of all RCTs are inherently reliable. First, the sample size must be large enough to identify outliers. Randomising only two classes or schools does not constitute a true RCT, as there may be pre-existing differences between the two groups. For example, differences in teacher effectiveness between the two classes or the fact that one school might be a high-performing institution with better resources can result in an unfair comparison, even if random allocation is used. Such a scenario can lead to false negatives or false positives. This issue was illustrated in my pilot study, which included only two classes. As the purpose of my pilot study was to rehearse the delivery of the intervention, trial the teaching and learning resources, and test the measuring instruments, as well as to practise the analysis, the sample size was not of primary importance.

A rigorous well-conducted RCT with a large sample using independent measures with low attrition can provide the best estimates of the impact of the intervention (Gorard, 2021). However, some concerns exist about the ethical issues of RCTs in the educational field. For instance, Oakley et al. (2003) argued that while RCTs are feasible in clinical settings, it is unethical to use them in education because students not receiving the treatment (in the control group) may be disadvantaged. This argument is naïve and flawed, as it assumes that the intervention is effective (Torgerson & Torgerson, 2001). The intervention may be harmful, so exposing all participants to the intervention without clear evidence of its efficacy is also unethical. A more sensible approach is to use a waiting-list design, where students in the control group can still receive the treatment after the RCT if the intervention is found to be effective. This way, no student is disadvantaged.

The primary research was a two-group cluster RCT where teachers were randomly allocated to either experimental or control conditions. The experimental group was taught using EnglishFusion CT lessons, while the control group continued with the regular English curriculum. Considering that the intervention was delivered in schools, it was not feasible to randomise students individually. This was to respect the existing teaching structure (Gorard, 2021), where students had already been assigned to different classes. Participating schools in the primary study did not require students to move classes to learn different subjects. Students from the same class were always taught together. If students were randomised individually, the regular teaching schedule would be disrupted, and more effort would be required to arrange the delivery of the intervention.

Randomisation at the school level was also not ideal because of the small number of schools involved ($n = 4$). This would result in only four cases, significantly reducing statistical power.

Additionally, as some teachers taught two classes, randomisation was conducted at the teacher level rather than the class level. Specifically, if classes were randomised instead of teachers, the same teacher might teach a control and an intervention class, thus introducing the risk of diffusions (Torgerson & Torgerson, 2001), where teachers might

unconsciously use EnglishFusion methods in the control class. Therefore, this study adopts a cluster, two-armed RCT where teachers, rather than classes, are randomised.

6.5 The sample

The sampling strategy

Schools were recruited using convenience sampling. The local Department of Education in one county in Sichuan province, China was initially contacted. The project and intervention were explained to department leaders, who then identified four schools they believed would benefit from the intervention. These four schools were purposefully selected for their heterogeneity in terms of school size and the various education levels they included.

To facilitate access to the selected schools and ensure the study was effectively managed, I specifically designed a workbook for school leaders and teachers, providing a detailed explanation of the research project. The workbook included an introduction, background information, and the purpose of the trial, alongside a clear outline of specific tasks and the research schedule. To increase their interest in participating, examples of how urban schools teach CT and highlights from recent conferences emphasising CT teaching were included. This helped contextualise the study's importance and relevance. Communications with school principals and administrative staff were then conducted to further clarify expectations, address concerns, and finalise consent. Without the help of the local department of education, it may be more challenging to recruit schools. It should be acknowledged that while participation was explicitly voluntary, with schools free to withdraw at any time, there is implicit authority pressure influencing their decisions. As a result, all four schools approached agreed to participate in the study.

Once the schools agreed to participate, all classes from Grade eight in the schools, including their English teachers, were included in the study. The rationale for using Grade eight rather than Grade seven students in this study is that Grade eight students were believed to have a sufficient level of English language skills. This is particularly relevant as students in village schools usually do not start learning English until they

enter secondary school. On the other hand, Grade nine students were avoided as this is the exam year when students prepare for high-stakes entrance exams.

Randomisation

As explained earlier, randomisation was conducted at the teacher level. Among the four schools, there were 21 English language teachers, forming 21 clusters in total. Teachers were assigned numbers based on the alphabetical order of their surname. Using a random number generator, teachers corresponding with the generated numbers were assigned to the experimental group, while the others were assigned to the control group. Table 6.3 shows the details of the teachers, schools, and the number of students in each group.

Table 6.3 Results of teacher randomisation

Experimental group			Control group		
Teachers (N=11)	Students (N=1027)	Schools	Teachers (N=10)	Students (N=1028)	Schools
Teacher 01	103	A	Teacher 04	107	A
Teacher 02	126	B	Teacher 08	118	B
Teacher 03	98	A	Teacher 09	108	A
Teacher 05	65	B	Teacher 10	109	A
Teacher 06	108	A	Teacher 11	72	D
Teacher 07	99	A	Teacher 12	108	A
Teacher 14	103	A	Teacher 13	128	B
Teacher 15	131	B	Teacher 19	70	B
Teacher 16	36	D	Teacher 20	104	A
Teacher 17	53	C	Teacher 21	104	C
Teacher 18	105	A	\	\	\

The randomisation resulted in 11 teachers with their students (N = 1,027) being assigned to the experimental group receiving EnglishFusion, while 10 teachers with their students (N = 1,028) were assigned to the control group, where they continued with their regular lessons as per normal. Experimental teachers were expected to infuse CT into their lessons and explicitly make the cultivation of CT a course objective. To

help avoid contamination, the experimental group was instructed to keep the intervention-related materials confidential.

Prior to randomisation all pupils in the study took a baseline test of CT skills and completed a questionnaire survey about their demographic background (see Appendix D1). Teachers and students were not informed of their group allocations until after they had completed the questionnaire and pre-test. This blinding was done to prevent any potential influence that knowledge of group allocation might have on students' attitudes towards the test or on teachers' unconscious bias towards one group over another.

For any trial, it is important to maintain a large sample size (Gorard, 2021; Torgerson & Torgerson, 2001). A large sample size increases the likelihood of obtaining accurate average values, helps identify outliers in the data, and provides smaller margins of error and lower standard deviations, leading to more trustworthy results. Additionally, it minimises the risk of reporting false positive (Type I error) or false negative (Type II error) findings. But how large is large?

Most researchers use what is commonly referred to as a “power calculation” to estimate the minimum sample size required for a study. However, as Gorard (2021) has demonstrated, this “power” calculation is problematic because it is based on significance tests that do not give us the answer that we want. That is, they do not tell us if there is a difference between groups. What the *p*-value tells us is: Assuming there is no difference between the groups, how likely are we to get the results that we do. This is not the same as: given the results, what is the probability that there is no difference between the groups – a very common misunderstanding.

In this study, I used Gorard's formula to estimate the minimum sample size based on the number of “counterfactual” cases needed to alter a finding (Gorard & Gorard, 2016). The number of counterfactual cases required to disturb a finding, or NNTD (Number Needed to Disturb), is calculated by multiplying the “effect” size by the number of cases in the smallest group in the comparison (i.e. the number of cases in either the control or treatment group, whichever is smaller). According to this sensitivity test, an NNTD of 45 can be considered a strong and secure finding. Assuming an “effect” size

of 0.2 (which is typical of educational interventions when there is an effect), a sample size of 225 per group ($45/0.2$) would be sufficiently sensitive to detect an effect. In this study, there were 1,027 pupils in the experimental group and 1,028 in the control group. The sample size is therefore large enough to safely suggest an effect size as small as 0.2.

6.6 Outcome measures

Primary outcomes

The primary outcome was CT skills. Altogether there were 15 questions about CT skills. CT skills were measured using modified versions of three standardised CT tests: the Cornell Critical Thinking Test Level X (Ennis et al., 2005a), the Halpern Critical Thinking Test (Halpern, 2005), and the Watson-Glaser Critical Thinking Appraisal (Watson & Glaser, 2012). The tests used in this primary study for measuring CT skills are presented in Appendix D.

The Cornell Critical Thinking Test Level X was chosen because it was designed for young students aged between four and fourteen in an educational context (Ennis et al., 2005b). This was appropriate for the pilot and main trial, where pupils were from lower secondary schools and for whom English was not their first language.

The Cornell Critical Thinking Test Level X assesses four sub-skills of CT: induction (hypothesis testing), credibility of sources and observation, deduction, and assumption identification. Other sub-skills, such as dealing with sensitivity to meaning and handling equivocation, are not included as they are deemed too complicated for young people (Ennis et al., 2005b). While this test uses a storytelling approach to keep students engaged, it is time-consuming as test takers need to read the background and plot of the story before answering the 71 test items. The suggested testing time is 50 minutes, which is relatively long for students.

For this trial, the number of items was reduced to 15, and the test duration was limited to no more than 30 minutes. This adjustment aimed to reduce test fatigue and boredom, which are likely if the test is too long. Additionally, it was important to consider that schools are generally reluctant to allocate excessive curriculum time for tests, preferring

to use this time for teaching. Therefore, not all items from the Cornell Critical Thinking Test Level X were used. Instead, a question bank was created, from which items appropriate for the students' age group and relevant to their context were selected.

The Halpern Critical Thinking Test presents 25 daily life scenarios, requiring close-ended answers and justifications for responses (Halpern, 2005). This test assesses different cognitive skills (Bridgeman & Moran, 1996) and reduces the likelihood of guessing compared to using only multiple-choice questions. It examines five sub-skills of CT: verbal reasoning skills, argument analysis, hypothesis testing, skills of using likelihood and uncertainty, and decision making and problem solving (Halpern, 2005; Ku et al., 2006). The full test was not adopted due to the additional time required for justifying answers and the assessment of writing skills, which were not the main focus of CT skills. Scoring constructed answers would also require pre-specified criteria and multiple raters for consistency, making it challenging for a large sample size. Therefore, questions from the Halpern Critical Thinking Test were added to the question bank for testing CT.

The most recent UK version of the **Watson-Glaser Critical Thinking Appraisal** (Watson & Glaser, 2012) was included more extensively in this study. This test, dating back to the 1920s, has undergone multiple refinements and is widely used in business recruitment and educational contexts (Watson & Glaser, 2012). The test measures five sub-scales of CT: arguments, assumptions, deductions, interpreting information, and inferences. Each section includes instructions explaining the sub-skill, what will be shown, and how to select answers. However, some instructions are too detailed for students and may include theoretical terms that could discourage them. Additionally, the requirement to assume the provided statements are true may not be met, as people tend to reject conclusions that do not align with their experience or knowledge (Evans, 2005). To address these issues, the instructions were modified to be concise and included only essential information understandable to English language learners.

The three established CT tests considered different subsets of CT skills, reflecting the various and debated definitions of CT. While there is some overlap, such as assumption and argument evaluation, it was not possible to include all aspects of CT. Therefore,

reductions were necessary. According to Ennis et al. (2005b), both skills of evaluation and understanding meanings should be included in a general CT aptitude test. To suit the age of the pupils and the secondary education context in China, five subsets of CT were assessed: evaluating arguments, identifying assumptions, making deductions, drawing inferences, and interpreting information. Each section included three question items, starting with the simplest to keep students engaged and increase their confidence when faced with more challenging questions. In total, there were 15 multiple-choice questions, which students were required to complete within 30 minutes to ensure they remained focused.

Specifically, the skill of **argument** evaluation was assessed in the first section. Students were presented with a question and needed to select the most relevant and reasonable argument. Here is an example of the item used in this section.

Figure 6.4 Sample item for argument evaluation

Should anonymous posting and commenting on the internet be banned?	<input type="checkbox"/>
A. Yes, this would reduce cyberbullying because perpetrators using their real identities would be held accountable.	
B. Yes, this would reduce cyberbullying because people would stop using the internet.	
C. No, because people should be free to comment on the internet.	
D. No, because posting images and comments on the internet does not harm anyone.	

The second section tested the ability to identify **assumptions**, where students were given a statement assumed to be true and asked to choose the hidden, established idea of each statement. An example of the question is shown as follows.

Figure 6.5 Sample item for assumption identification

Statement: I saw two people across the road wearing hats. The shorter of the two was a female. I say this because I saw her long hair when she removed her hat.

The assumption is that:

A. Only females have long hair.
B. All females are short.
C. All females have long hair.

The third section focused on drawing **deductions**, requiring students to read a statement and come to a conclusion based solely on the provided information, without using their general knowledge or personal opinions.

Figure 6.6 Sample item for deduction skill

Statement: Only technological companies are listed on the OTX (a computer security platform). No technological company remains unstable for a long time.

The conclusion is that:

A. If one company is not unstable for a long time, it will be listed on the OTX.
B. If one company is listed on the OTX, it will be unstable for a long time.
C. If one company is listed on the OTX, it will not be unstable for a long time.

The fourth section remained the same as the Watson-Glaser Critical Thinking Appraisal, presenting a short paragraph describing an issue and three subsequent questions. Students had to evaluate the trustworthiness of **inferences**, choosing from five options: true, probably true, more information required, probably false, and false.

Figure 6.7 Sample item for inferential skill

Statement: Two hundred students in their early teens voluntarily attended a recent weekend student conference in London, England. At this conference, the topics of race equality and ways of achieving world peace were discussed, since these were the problems the students selected as being most important in today's world.

10. Inference: These students came from all parts of the UK.

Based on the statement, this inference is:

- A. True
- B. Probably true
- C. More information required
- D. Probably false
- E. False

The final section measured the ability to comprehend and **interpret** information, where students needed to understand the text well to choose the correct answer.

Figure 6.8 Sample item for interpreting information

Text: A recent report in a magazine for parents and teachers showed that adolescents who smoke cigarettes also tend to get low grades in school. As the number of cigarettes smoked increased, students' grades decreased. One suggestion made in this report was that we could improve students' grades by preventing adolescents from smoking.

The conclusion is that:

- A. The suggestion is supported because the research found that smoking causes grades to decrease.
- B. The suggestion is supported because the research found that reducing smoking can improve grades.
- C. The suggestion is not supported because the research does not show that smoking causes grades to fall.

The modified CT tests employed in this study were administered in paper and pencil format, familiar to Chinese students in their regular school learning. Since the original language of the three independent CT tests was English and secondary school children in China were not generally proficient in English, a Chinese version was provided. This was reasonable as the test aimed to assess CT skills rather than English language skills. To ensure accurate translation, a reverse translation method was used. Additionally, the length of alternatives for each question was kept similar to reduce the likelihood of guessing based on option length. In many circumstances, the correct option includes more details and might be longer (Tarrant & Ware, 2008). To accommodate students who preferred to read the original content, an English version was also provided. The scoring method was “right only”, meaning students could only earn points for selecting the correct option. The maximum score for each test was 15. To avoid familiarity with the test items, students were tested after the intervention with a post-CT test that measured the same CT sub-skills using different questions (see Appendix D2).

These modified CT tests have been piloted. They were deemed appropriate in terms of difficulty level, and the testing time was also suitable, with all students able to complete the questions within 30 minutes. However, there were several improvements to the main study. Test items translations were fine-tuned to ensure there were no omissions or misunderstandings. Some questions from the pre-test were switched to the post-test, and vice versa, to balance the difficulty level between the two CT tests. The sequence of items within each section was also considered, with students initially encountering easier questions. In addition, one item that asked students to deduce the likelihood of an incident had an approximately 97% correct rate. This was not effective in distinguishing between students with different levels of deduction skills. This item was thus replaced by a new deductive question (Koretz, 2006).

The pilot study also informed the circumstances of students doing these CT tests. The pilot teacher noticed that around 20% to 40% of the students were not particularly serious about taking the CT tests. Their different attitudes towards the tests might influence the test results. To address this, teachers provided assistance during test administration and monitoring in the main trial. I brought CT tests to each school and

explained what they were expected to do. This helped students understand the importance of the CT tests and remain focused.

The CT tests were administered to all the Grade 8 students in the four participating schools. The pre-test was completed on 17th February, and the post-CT test was completed on 7th June, after the three-month intervention. Teachers helped organise and maintain the testing environment but were not informed of the test content in advance to avoid any possibility of teaching to the test.

To minimise the risk of diffusion of test questions, students from the same school took the CT test simultaneously. Schools C and D, having a smaller number of students, were able to arrange for all students to take the test in a large conference room. However, this was not possible in the larger Schools A and B, where students took the tests in their usual classrooms. A few students were absent on the day of the test. These students were followed up and tested when they returned to school.

Secondary outcomes

The secondary outcome was academic achievement, measured by county-level final examinations. As Chinese, Maths, and English are the three core subjects in secondary education in China, the academic scores for these subjects were collected as indicators of academic attainment.

In China, secondary school students are generally required to take a county or district examination each term. This summative test assesses how much students have learned during the term. Teachers from that area are randomly and blindly assigned to mark the tests, which helps to avoid bias. Since the four schools involved in this study were from the same area, students took the same test and were marked according to the same criteria, ensuring consistency across schools.

The pre-academic scores were the final exam results of the previous term, indicating the students' most recent academic performance. Due to the COVID-19 school closures, students in this area took the final test at the start of the term instead of at the end of the previous term. After the three-month intervention, students sat for the final

examinations on the 3rd and 4th of July, and these results were regarded as the post-academic scores.

Notably, in the pilot study, student post academic attainment scores were not collected because the COVID-19 pandemic interrupted usual school learning, and students from the pilot school did not attend the final examinations at the end of the term. However, it was suggested that some students might forget or even fabricate their scores in the pre academic score collection. Therefore, to increase data accuracy in the main trial, the scores were obtained from teachers rather than students.

Other data

Teacher questionnaire (see Appendix E)

Teachers play a crucial role in delivering CT lessons and improving students' CT skills (Choy & Cheah, 2009). They need to guide student thinking by giving prompts rather than directly providing answers. This can be challenging for some teachers who consider themselves the authority of knowledge and neglect their students' voices. Moreover, the willingness and ability of teachers to deliver educational interventions influence the fidelity of implementation (Stein et al., 2008). For these reasons, a teacher questionnaire was designed to collect teachers' demographic characteristics, critical awareness and views on CT teaching to see if these teacher factors are associated with student outcomes. It also aims to collect information for process evaluation.

The front page of the teacher questionnaire was titled "Infusing Critical Thinking in English Lessons". Participants were reminded of the upcoming intervention, and the purpose of the questionnaire was briefly explained. The estimated time to complete it (about five minutes) was provided to give teachers a general impression of the workload, allowing them to plan their time accordingly. While the teacher questionnaire was in English, teachers were informed that they could use either English or Chinese, whichever suited them better. This was to reduce their language or translation burden and ensure the accuracy of their responses. Anonymity and confidentiality were assured, and reasons for requiring their names were explained (for identification to match their pre- and post-results). My contact details were also provided in case of any queries or unforeseen problems.

The teacher questionnaire consists of two sections. **Section A** asks four questions. Question 1 measures teacher's critical awareness, Question 2 asks about teachers' attitudes towards CT teaching in school, Question 3 measures the frequency of CT-related activities in usual English lessons, and Question 4 is about teachers' perceptions of barriers to teaching and learning CT. Some question items were reversed to reduce response bias (Paulhus, 1991). The 11-point Likert scale was used in all four questions to measure responses, with 10 indicating strong agreement and 0 indicating no agreement at all (Harpe, 2015; Leung, 2011). This approach was chosen to capture more nuanced differences in responses. The conventional 5-point Likert scale was avoided because the intervals between each point are not equal. By using more points, the scale reduces skewness and better approximates an interval scale (Leung, 2011; Wu & Leung, 2017). This allows responses to be treated as real numbers and enables the calculation of effect sizes using mean scores.

Question 1 asked teachers to judge the trustworthiness of research findings in newspapers or magazines. Five statements on the currency (publication date), reputation of journals, authority of authors, inclusion of standardised tests, and sample size were provided, and teachers indicated their level of agreement by ticking the 11-point Likert scale.

Originally in the pilot study, this question assessed teachers' knowledge of CT. However, knowing the meaning of "critical" in CT does not equate to possessing CT skills. Therefore, in the main trial, teachers' CT skills were directly assessed. This is important as their CT skills play an important role in cultivating students' CT (Wang & Jia, 2023). In teaching CT, teachers are required to model the thinking process, clearly explain complex issues, and provide concrete examples (Lai, 2011). If their CT skills are poor, students may find it difficult to understand the CT lesson content.

Question 2 asked teachers to indicate the level of agreement with four statements about the importance and relevance of teaching CT in schools, particularly within the English curriculum and teacher training, using a scale from 0 to 10.

This question was modified based on Stapleton's (2011) survey of teachers' attitudes towards CT. The medium sample size ($N = 72$) and the high Cronbach alpha (0.703) indicated good reliability of this instrument. However, different research contexts and aims necessitated revisions to the existing instrument (Oppenheim, 2000). While the original survey contained eight statements, only relevant statements were selected in this trial. Additionally, the previous instrument was designed for teachers in various subjects (e.g. science and mathematics, humanities, and physical education), whereas this research focused exclusively on English. Therefore, statements were made more specific to the relevance and importance of CT in English. The original eight-item Likert scale was transformed into an 11-point Likert scale to increase response sensitivity and maintain consistency with other questions in this section.

Question 3 asked teachers to rate the frequency with which students in English lessons engage in specific CT-related activities, including memorising facts, explaining answers, applying knowledge, thinking of alternative explanations, questioning information, and creating new ideas, on a scale from 0 (never) to 10 (always).

These activities were based on Bloom's taxonomy (Krathwohl, 2002). Keywords were appropriately phrased to help participants achieve consistent meanings for each statement (Boynton & Greenhalgh, 2004). For example, the statement "justify a stand or evaluate a decision" in the pilot was rephrased to "question the trustworthiness of information received" for clarity. A Chinese translation of the word "trustworthiness" was provided to facilitate comprehension.

Question 4 was about barriers to teaching critical thinking in schools. These factors were identified in existing literature, which included: the debated definition of CT (Fisher, 2011), a lack of sufficient background knowledge (McPeck, 1981), examination-oriented teaching (Jiang, 2013), large class sizes (Guo & O'Sullivan, 2012), respecting authority figures (Paton, 2005), and insufficient training in teaching CT (Snyder & Snyder, 2008).

The display of each item was intentionally mixed to prevent teachers from identifying any pattern or tendency, ensuring that they focused on reflecting their own opinions on

the obstacles to teaching CT. For each barrier, teachers indicated their degree of agreement, rating from 0 (do not agree at all) to 10 (completely agree).

Recognising that close-ended questions might not cover all potential challenges, a free text space was provided after these listed obstacles (Boynton & Greenhalgh, 2004). This allowed teachers the opportunity to write down any additional perceived factors with explanations.

Section B is about teachers' demographic factors, including age, birth sex, educational background, and work experience, were collected. There were five questions in this section. These questions were asked to see if such demographic characteristics have any bearing on teachers' attitudes towards teaching CT and how they deliver CT in the classroom. Such information may be sensitive. Asking such questions at the start of the questionnaire may put some people off. Putting these questions at the end ensures that questions about teachers' attitudes toward CT teaching, which are the substantive questions, are given priority.

Question 5 asked for teachers' age. A free text box was provided so they could write down their age in years. Since these participants were unlikely to be of the same age, their age was requested in years only.

Question 6 asked for teachers' birth sex. Three options were provided for selection: male, female and prefer not to say. Since terms like "sex" and "gender" are often used interchangeably in research (Doyal, 2003), the word "birth sex" was used consistently to avoid confusion.

Question 7 was about teachers' educational experience. Three sub-questions were asked successively, each addressing independent aspects of teachers' educational backgrounds. First, teachers indicated whether they had attended a normal university via a binary yes or no question. Normal universities are typically teacher training institutions, so it is assumed that graduates from these institutions have pedagogical training.

Second, they were asked about their highest educational qualifications: whether they had an undergraduate degree (equivalent to a bachelor's degree in the UK), Master's and doctorate degrees. Considering that teachers may experience different educational systems and that it is not possible to cover all degrees, an option labelled "Other" was provided. Teachers choosing this option could specify their highest academic qualification in a free text box.

Third, they were asked a binary yes or no question about whether they attended an overseas institution for their degree. If they chose 'no,' this implied that their educational experience primarily took place in China. If teachers indicated that they had studied abroad, they needed to clarify whether it was in an English-speaking area. This question was asked in detail because CT is not always a primary focus in different educational settings, and the emphasis on CT development varies across policies. For example, South Korea did not incorporate CT education (McGuire, 2007) until recent decades, and it still lacks concrete CT teaching approaches in various curricula from elementary schools to universities (Ro, 2023). However, CT is often explicitly mentioned in English-speaking educational institutions (Andrews, 2007; Greenholtz, 2003). If teachers pursued a degree in English-speaking countries, they were likely exposed to CT. In this case, they would understand CT lesson content more easily and be more willing to deliver the CT lessons, although it remained uncertain whether they would teach as planned.

Question 8 asked for the number of years teaching English in secondary schools, which could be recognised as an indicator of work experience. Teachers wrote down the exact number of years in a free text box. Generally, those with longer teaching careers are more likely to have richer work experience.

At the end of the teacher questionnaire, **Question 9** provides a free text box for teachers to indicate any other relevant experience. Sufficient space was provided, so teachers would not be hindered by a word limit. They could share their experiences or other issues regardless of length.

To track any changes in teachers' CT awareness and attitudes towards CT teaching, a self-report questionnaire with minor revisions was administered to teachers at the end of the trial. **The post-questionnaire** had two refinements. First, the instruction on the front page was updated to be consistent with the research procedure. Teachers were informed that their names were required to match those in the "previous" questionnaire, not the "later" one. Second, as their demographic data had already been collected in the pre-questionnaire and was unlikely to change after the intervention, Section B was removed.

The pre-teacher questionnaire was sent to all 21 teachers before the random allocation. After the three-month implementation of the intervention, teachers were asked to complete a post-questionnaire. These questionnaires were distributed via an online platform, which was more convenient for teachers compared to the paper-and-pencil method. While using online survey software can save time and reduce data entry errors (Fife-Schaw, 2001), the questionnaires in this study were formatted as Word documents. This format was chosen because the questionnaire included multiple ways of indicating answers, such as ticking, selecting, and writing, which would be difficult to incorporate into an online questionnaire.

Notably, the Word document format cannot guarantee that participants answer all questions before moving on to the next sections (Fife-Schaw, 2001), so some questions were left blank in the pre-questionnaire. In this case, teachers were contacted to supply the missing information. Additionally, some teachers' answers showed the same pattern (e.g. all choosing 5 or 10), which was less likely to be genuine. To increase the trustworthiness of responses, these questionnaires were returned to the respective teachers, who were asked to provide their true thoughts and base their answers on real circumstances. For the post-questionnaire, there were no missing responses or patterned answers. All teachers completed the post-questionnaire once, and no reworking was necessary.

Student questionnaire (see Appendix D1)

The student questionnaire collects information about student background and demographic characteristics: sex, ethnicity, age, socioeconomic status (SES) and parental involvement in their education. There are five questions in this section.

The reason for collecting students' background information via a questionnaire was to establish equivalence between the two groups prior to the trial to ensure they were balanced. These factors are also known to be associated with students' CT performance and academic outcomes (Deal & Pittman, 2009; McCutcheon, Hanson, Apperson & Wynn, 1992). If the two groups are not balanced, it means that one group already had an advantage at the outset. To attribute any effect to the intervention, it is necessary to control for these differences. Additionally, sub-group analyses categorised by students' demographic variables were also performed.

The student questionnaire forms Part II of the pre-test for CT (see Appendix D1). It collects information about students. CT test questions were asked in the first part (Part I). This was to help students focus more on answering the CT questions, which required a higher level of cognitive skills. Questions about student background information were asked following the test as these were deemed easier for students to complete and even if students had missed these questions, substantive questions about their CT would have been collected. They were also assured that their responses would be kept confidential. Their teachers and schools would not know their responses.

Following the 15 CT test questions, students were asked about their demographic data. **Question 16** is about students' sex. This was a binary question: male or female. The term "birth sex" was used to emphasise biological characteristics that are more distinguishable and stable. This is a common method of collecting birth sex information in China, and students are familiar with it.

Question 17 was a close-ended question that addressed students' ethnicity. This was also a binary question: Han or minority. Students could indicate their ethnicity based on their identification card, which is unique to each citizen in China.

As students were similar in age (being in Grade 8), the specific date of birth was asked via **Question 18** to differentiate older from younger students. A free text box was provided where students were instructed to write down their birth date in numbers (e.g. birth year, month, and day). This format aligns with the date format in China, and students are accustomed to displaying their birth dates this way. To ensure clarity, an example was provided. The age was calculated by subtracting the date of birth from the date of the pre-test of CT.

Question 19 collected information on students' social-economic background. As it is not permitted in China to collect information about schoolchildren's household income, parental occupation and educational level directly, students were asked about their household possessions as a proxy measure of SES. This policy is intended to prevent students from being disadvantaged or privileged due to their socio-economic background. This posed a challenge for the study because students' SES is correlated with their academic performance (Gorard & See, 2013; Liu & Lu, 2008).

To address this issue, the families' economic situation was indirectly assessed through questions about students' household possessions. Ten items were selected from the OECD's survey on social and emotional skills (2021) based on their relevance to the Chinese context. These included personal rooms, study desks, computers for homework, Wi-Fi, bookshelves, classic literature, books of poetry, works of art, books on art, music or design and musical instruments.

For each item, students had to tick either 'yes' or 'no' to indicate whether they had the object in their home. The 'no' option was included to clarify responses, as a missing tick could either mean the object was not present or the question was overlooked. Additionally, students were instructed not to tick both options, as doing so would render the response invalid.

Question 20 was about parental involvement in children's education. Students were given a list of five activities that they might do with their parents. These included discussing school performance with children, helping children with homework, discussing political or social issues, going to a library or bookstore together and talking

about children's reading. Students were asked to indicate the frequency of their parents' involvement in each activity over the last academic year, rating from Never (0) to All the time (10).

This question was asked because it is widely believed that parental involvement in children's education influences their educational outcomes (Đurišić & Bunijevac, 2017), and perhaps their CT as well (Spence, 2012). These five activities were adapted from the OECD PISA questionnaire (OECD, 2018b). Unlike the original assessment, which was designed for parents, this study asked students to report on parental involvement in their education. Hence, the phrases were modified accordingly. Additionally, specific activities closely related to students' education were selected. For example, items about students' school performance and homework were included, while the frequency of parents eating the main meal with the child was not considered, as it was deemed less relevant to educational participation. Moreover, the original scale was less sensitive, so an 11-point scale was employed.

In the pilot study, the student questionnaire asked students to provide their exam scores because the pilot teacher did not have access to students' scores in Chinese and Maths. However, some students were unable to complete this question as they had forgotten their exact scores. This question was therefore removed in the main study and the data on academic scores were directly obtained from the local Department of Education in the main trial. It was also found in the pilot that some students did not read the questionnaire instructions carefully. In the main study, key guidance and questions were highlighted in the main trial, and clear instructions were emphasised before administering the questionnaire.

6.7 Analyses

This section outlines the methods of data analysis used in the primary research. As McCoy (2017) notes, different methods of data analysis can yield different results. Using an incorrect method in an intervention study can lead to biased evaluations of the intervention. Furthermore, data dredging bias can occur when data is intentionally probed and analysed using multiple approaches (Erasmus, Holman, & Ioannidis, 2022). Several decisions must be made during the data analysis process, such as handling

missing values and measuring differences between groups. To avoid data dredging and ensure transparency in data analysis, the details and justifications for the analyses are presented as follows.

Impact evaluation

The intervention's impact was measured by the differences in CT skill test gain scores. The difference between groups was estimated using Hedge's g effect size (ES), which indicates the magnitude of differences between the two groups (Sullivan & Feinn, 2012). The ES is calculated by dividing the mean difference between the experimental and control groups by the overall standard deviation (SD).

In addition to the overall CT gain scores, scores for sub-skills of CT (i.e. argument, assumption, deduction, inference, and interpretation) were also obtained. To examine the intervention's impact in greater detail, the ES of gain scores for each CT sub-skill was calculated.

Sub-group analyses of CT gain score differences, categorised by demographic variables of students including age, birth sex, ethnicity, SES, parental involvement in education, schools, and teachers, were performed using ES. This approach was similar to the sub-group analyses on academic attainment.

It is important to note that significance tests, such as p -values and t -tests, were not considered in this study. Although these tests are widely used to justify the statistical significance of substantive results, their application in real-life studies is problematic. On the one hand, their practical application is overly idealistic (Gorard, 2021). Employing a significance test requires cases to be selected completely randomly (Colquhoun, 2014; Gorard, 2021). However, this is almost impossible and unrealistic in real-life research because random cases imply no missing data, no errors in measurement, and no data entry mistakes (Carver, 1978; Gorard, 2021). While this study did randomise students into two cohorts, it was a cluster randomisation at the teacher level. Additionally, there were missing cases and values in this three-month intervention study. Therefore, since the study did not satisfy the logical premise for using significance tests, it is reasonable to ignore them. Moreover, significance tests do

not provide the actual probability of results arising by chance (Gorard, 2021). To avoid misleading findings and poor conclusions (Ioannidis, 2005; Tarran, 2019), this study did not use significance tests.

Multiple linear regression analyses

To consider further whether any differences in CT after the intervention could be attributed to the intervention itself, a multiple linear regression analysis was performed. This analysis considers the interaction or correlation between two or more variables (Gorard, 2021). By controlling for some of the background factors of teachers and students, it is possible to determine the relative contribution of each explanatory factor in explaining the variance in the outcomes.

For the primary outcome (i.e. CT skills), the post-CT score was selected as the dependent variable, and other independent variables were entered in blocks, sequentially in chronological order. The first block included student demographic factors such as age, birth sex, ethnicity, household items, and parental involvement in education. These were included first as these are factors that are not malleable. The second block included the pre-test score of CT. Prior academic attainment was included in the next block. Schools and teacher factors were included in the fourth block, and the final block considered whether students were in the treatment group or not to see how much more participation in the intervention could explain children's performance in CT.

For the secondary outcome (i.e. academic attainment), the post-academic score was used as the dependent variable, with other factors entered accordingly: background, pre-test academic scores, pre-test CT scores, schools and teachers and the membership of intervention.

To identify the best predictor of dependent variables (i.e. post CT scores and post academic scores), a forward selection method was employed. This approach allows for a simpler presentation with fewer predictors, while still accurately predicting the outcome (Gorard, 2021). Since there were several predictors in this model, the adjusted R-squared was calculated to produce a predictive model. The increase in the percentage

of variance explained at each step shows how much more these variables add to explaining children's CT scores after accounting for the previous set of predictors.

Sensitivity analysis

To evaluate the robustness of the results, a sensitivity analysis was conducted (Thabane et al., 2013). As with many trial studies, there were missing cases in this study (Gorard, 2021). To make the attrition explicit, the number of cases recruited, allocated and missing, were recorded and reported in the Consolidated Standards of Reporting Trials (CONSORT) flow diagram (Moher et al., 2010, see Chapter 9).

It is a common practice to use existing data to substitute for missing data under the assumption that the missing data are random. However, research indicates that missing cases usually do not occur randomly (Gorard, 2021). Students missing tests are likely to be those excluded/suspended or long-term sick, school refusers or have special learning difficulties. Excluding them from the final analyses could skew the results and inflate the effects (Dumville, Torgerson, & Hewitt, 2006). To address this issue, pre-test CT scores of missing students were compared to those of completed cases. The comparison could assess whether students who were lost to follow-up and others shared similar characteristics.

Another solution to dealing with attrition, as proposed by Gorard & Gorard (2016) is to report any missing data and compare the level of missing data to the number of hypothetical counterfactual cases needed to disturb the finding or NNTD (Number Needed to Disturb). It estimates what would have happened if the outcomes scores of those missing test scores were included. Since it is not possible to know what test scores those who have not taken the post-test would be, NNTD calculates the number of counterfactual cases with the opposite results needed to be added to the smallest group in the comparison before the 'effect' disappears.

NNTD considers whether the results would be altered if all missing cases were regarded as counterfactual ones (Gorard, 2021). It is calculated by multiplying the effect size (ES) by the number of cases in the smallest group in the comparison (i.e. the number of cases from the experimental group in this study). If the NNTD is larger than the number of

missing cases, it can be concluded that the results are robust and stable. However, if NNTD is smaller than the number of missing cases, this means that the results are unstable. The effects would be zero or reversed if the scores of the missing were included.

6.8 Process evaluation

A process evaluation is a crucial component of all experiments, yet it is often overlooked. It provides valuable additional information about the trial, including the fidelity of the intervention's implementation (i.e. whether the intervention was delivered as prescribed), as well as the challenges and barriers to its implementation. Process evaluation can help identify the reasons why the intervention is effective or not, the mechanisms driving changes in the outcome (Bugge, 2024), and whether any modifications are necessary (e.g. whether the dosage is adequate).

For this reason, a process evaluation was conducted alongside the trial (Gorard et al., 2017; Oakley et al., 2006). Classroom observations were made to observe the delivery of the intervention. This was to determine fidelity to treatment (i.e. whether teachers adhered to the lesson plans) and to see if there was any diffusion (i.e. control students being taught using the CT materials). In addition, interviews were carried out with teachers and students. Lesson observations substantiated and corroborated teachers' self-reports about CT activities in their classrooms. Both field notes and transcripts of interviews help to identify common themes such as conditions for and challenges to successful implementation.

Class observation

The observation of experimental classes examined the implementation of CT lessons. It helped to verify whether teachers were able to deliver EnglishFusion as they had been trained to do. If students were taught as planned, their reactions and answers were noted to determine if there was any evidence of progress in CT. Any departure from the planned lessons was noted. For example, instances where teachers did most of the explanation giving students little opportunity to discuss, or directly gave answers, and moved through questions quickly, leaving students with insufficient time to think and

discuss, were recorded. Barriers or challenges to the implementation, such as inadequate English language ability and lack of time, were also identified.

To check for the potential for diffusion where control teachers may be using strategies similar to CT infusion, observations were also made of control classes. This is important as it can help explain the results.

In the first stage, seven CT lessons were observed on-site. Each teacher's CT lesson was observed and audio-recorded once a week. In total, 77 CT lessons (11 teachers *7 lessons) were documented, and audio recorded with permission from the teacher. To avoid disrupting the regular teaching schedules that had been arranged by the schools at the beginning of the term, teachers were asked to decide on an appropriate timeslot for observation. Lesson observations were carried out discreetly with me sitting at the back of the classroom, taking field notes. Appendix F is an example of lesson observation notes.

In the second stage, observations were made to see how teachers delivered the six CT tasks in their regular lessons. As it did not take a full lesson period to complete a CT task, and experimental teachers found it difficult to schedule the timetable for the on-site observation, they were asked to record their teaching of CT tasks and send the videos to me.

Interviews

As an important part of the process evaluation, interview data provided detailed insights into the implementation of the intervention. Semi-structured interviews with experimental teachers were conducted at the end of the study. Depending on the availability of teachers, those from the same school formed a teacher group and participated in a group interview (Schools A & B). One-to-one interviews were conducted in schools where there was only one experimental teacher (Schools C & D).

Each interview lasted between 20 and 35 minutes. Interviews were held in a quiet location in the schools. Teachers were asked what they thought of the lessons, and the activities. The interview questions were:

- What do you think of these CT lessons so far? What do you like or dislike about these lessons?
- How do you find the examples or the activities? Are they at a suitable level of difficulty for students?
- What do you think of the examples or the activities in terms of interest?
- What do you think of the level of the language? Is it simple, appropriate, or difficult for students?
- How do you think these lessons can be integrated with the current English curriculum?
- Tell me about students' interactions with the materials (e.g. student handouts, slides, etc).
- How useful do you find these lessons in terms of improving students' thinking skills and English learning?
- Have you ever taught students critical thinking? If so, what is the difference between your previous teaching and this one? If not, how is our course different from your expected thinking lessons?
- What aspects of the lessons do you think could be improved?

At the end of the semi-structured interviews, teachers had the opportunity to share any other thoughts that were not previously mentioned. It is notable that the specific interview questions varied. For instance, if teachers had already discussed the impact of CT lessons on students' English learning, the relevant question was not repeated. This means that the semi-structured interviews with teachers were flexible. Moreover, all interviews were conducted in Chinese, the native language of the participants. Using their native language allowed students and teachers to express their ideas clearly and accurately (Cortazzi, Pilcher, & Jin, 2011). The interview transcripts were later translated into English for further analysis.

Semi-structured group interviews with students from the experimental classes were conducted after the intervention. Three or four students per class participated in the interviews. Some students volunteered to be interviewed, while others were selected by teachers. Teachers had different selection criteria. Some teachers preferred students

who were positive about the intervention while some chose a mix of high, medium, and low performing students to get a more balanced view.

The interviews with students in the experimental group were conducted over a two-week period, from the end of May to the beginning of June. This relatively short delay between the end of the intervention and the interviews minimised the risk of affecting participants' recall of events. The number of students attending each interview ranged from three to eight, with the duration varying from approximately 24 to 51 minutes. Similar to the teacher interviews, the student interviews were held in a quiet space within the school to ensure that students felt secure and comfortable expressing their thoughts.

Before the interview, students were informed about the purpose of the interview and the confidentiality of their responses. Students were then asked about their experiences with the CT lessons, including the difficulty and interest of the examples or exercises, the impact on their English learning, thinking skills, and daily life, their previous experience with thinking classes, and any suggestions they had for improving the CT lesson content. The interviews were flexible, with additional specific questions asked based on students' responses to clarify or further explain their answers.

The list of interview questions is as follows:

- What do you think of these CT lessons so far? What do you like or dislike about these lessons?
- What do you think about the examples or activities in terms of difficulty? (Is the content new to you? Is there anything you do not understand? Which lesson is the most difficult? Is it the language, the content, or the way of teaching that makes it difficult?)
- What do you think about the examples or activities in terms of interest? (Can these examples or activities attract you and lead to an active discussion?)
- Do you find these lessons influence your English learning? If so, how? Can you give some examples?
- Do you find these lessons influence your thinking?

- How do you use what has been learned from the infusion lessons in your daily life? (Can you give an example?)
- Have you ever attended courses that help to improve thinking skills before? If so, what is the difference between your previous learning and this one? If not, how is our course different from what you imagined thinking lessons to be?
- How do you think the infusion lessons could be better? (e.g. the content, class atmosphere, teachers' guidance, course duration, learning materials, etc.)
- Will you recommend this course to your friends who have not learned critical thinking skills before? Why?
- Is there anything else you would like to talk about?

6.9 Ethics

Ethical approval for the pilot study was obtained on 20th July 2022 (see Appendix G1). Before the primary research started, ethical approval was sought from the Durham University School of Education Ethics Committee and approved on 14th November 2022 (see Appendix G2). Once the schools were recruited, teachers and students received documents such as a debriefing sheet, privacy notice, and information sheet. All students and teachers were informed about the research aims, as well as the methods of data security and storage. Since all students were under 18 years old, opt-out parental consent for participation in the study and the use and analysis of data were collected prior to the study. Opt-out consent was deemed appropriate as the intervention was conducted as part of the student's regular lesson. Parents who do not want their children's data to be used for the study would opt out. In any case, no parents opted out. All parents agreed for their children to be involved in the study. As the intervention was treated as a normal part of school teaching, consent from school leaders was also sought.

Students were assured that their responses to the questionnaire survey would be confidential. Although they were asked for their names, it was emphasised that this was for matching purposes (to match their pre-test answers to their post-test). Once pre- and post-test data have been matched, students' names were removed, and they were identified by a numeric code from 1 to 2055.

Before every class observation, students and teachers were assured that their responses would not be attributed directly to them. They will not be named in any report. All their responses will be anonymised. Students were allowed to speak freely about their experiences and views regarding the CT lessons, and there were no right or wrong answers. They were also assured that after the intervention, EnglishFusion materials would be packaged and distributed to the teachers and classes to ensure that those in the control group were not disadvantaged.

6.10 Chapter summary

This chapter clarifies and justifies the major components of conducting an RCT, such as randomisation, cases, and intervention. The cluster randomisation and large sample size help improve the quality of this primary study. With appropriate training in CT course design, I developed the three-month EnglishFusion intervention for Chinese students. EnglishFusion is described in detail and thus, allows for future replication. If it was found to be effective, these materials could be used in future training and teaching.

As an important part of the primary research, the pilot study was essential to test the intervention materials, lesson activities and questionnaire instruments. Corresponding changes were made in light of the pilot study. It also provides an opportunity to do teacher training, class observations and interviews. This helped plan for the main study and pre-empt any issues that might arise.

It also describes the methods of collecting and analysing different types of data. Decisions were seriously made and justified. The process evaluation is incorporated to offer additional insights into the delivery of the intervention within the research context.

Section III Results of the systematic review

This section consists of one chapter. Chapter 7 presents the results of the systematic review, focusing on the following research question: What is the evidence on Chinese students' critical thinking compared with students of other nationalities?

Specifically, it addresses these three sub-questions:

1. How do Chinese students' critical thinking skills compare with those of other nationalities?
2. How do Chinese students' critical thinking dispositions compare with those of other nationalities?
3. How do Chinese students' critical thinking styles compare with those of other nationalities?

This section outlines the results of the various stages of the review, from database searches to screening and synthesis. It details the number of studies screened at different stages and provides a summary of the quality of the included studies.

Chapter 7 Results of the structured review

To establish the claims that Chinese students are less critically aware than students of other nationalities, a systematic review was conducted. It addresses the research question: What is the evidence on Chinese students' critical thinking compared with students of other nationalities? A search of seven social science databases was carried out to identify relevant literature. The strength of the evidence of included studies was examined before drawing conclusions.

7.1 Results of the search

The initial search of seven electronic databases found 1,481 studies. Of these, 735 were duplicates and were removed. After applying the inclusion and exclusion criteria, only 13 reports (out of the remaining 746) were retained for full-text retrieval. The full text of two of these reports was not accessible. Thus, 733 studies were excluded for the following reasons:

- 237 studies failed to compare the critical thinking (CT) of Chinese students with those from other nationalities.
- 233 studies were not about the CT of Chinese students.
- 128 studies did not measure CT using validated or standardised tests (e.g. self-reported surveys).
- 106 studies did not have measurable outcomes.
- 18 studies were not empirical.
- 10 studies focused on item reliability and validity of CT assessments.
- 1 study was written in Russian.

An additional 1,471 studies were identified through manual searches. Screening by titles and abstracts removed 1,359 records that did not meet the inclusion and exclusion criteria. This process retained 112 reports for full-text screening. Of these, only 10 were deemed relevant to the research question and met the inclusion and exclusion criteria. The other 102 reports were excluded for the following reasons:

- 39 were duplicates.
- 29 were not about the CT of Chinese students.

- 17 did not measure the CT of Chinese students.
- 14 did not compare Chinese students with other nationalities.
- 1 was not empirical.
- 1 was written in Korean.
- 1 was not accessible.

Including both the online database and manual searches, a total of 21 reports deemed relevant to the research questions and meeting the inclusion and exclusion criteria were retained. However, these included some of the same studies that were there reported in multiple outlets such as dissertations and journal articles (e.g. Dennett, 2014; Dennett & DeDonno, 2021). Excluding these, only 15 unique studies were kept, and their findings were synthesised.

The PRISMA flow diagram (see Figure 7.1) tracks the review process from the identification of studies through online and manual searches to screening and analysis.

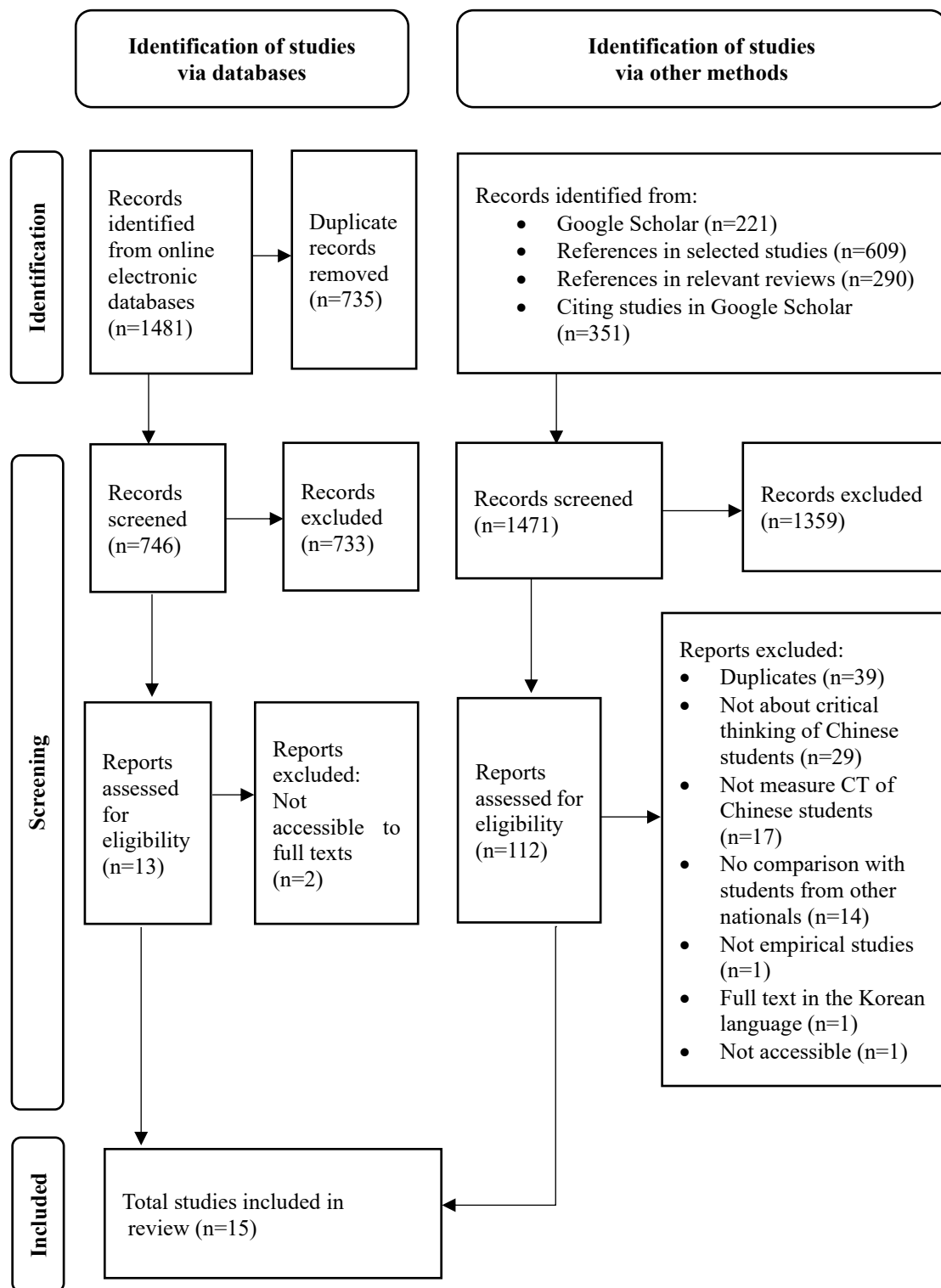


Figure 7.1 PRISMA flow diagram

7.2 Strength of evidence of included studies

Each of the 15 included studies were then assessed for their strength of evidence (see Table 7.1). No study was rated 4* (the highest level of credible evidence). Only one study was rated 3* and three studies received a rating of 2*. The majority (11 of the 15) were rated 1*.

Table 7.1 Summary of strength of evidence for all studies (N = 15)

Strength of evidence	CT skills (n = 8)	CT dispositions (n = 6)	CT styles (n = 1)
4*	-	-	-
3*	1	-	-
2*	2	-	1
1*	5	6	-

This indicates that the existing research evidence in this area is generally weak. There is therefore no strong evidence to suggest that Chinese students are any less critically aware or skilled than other students. Most studies had serious methodological defects, such as having small samples (e.g. Dong, Li, & Liu, 2010; Liu, 2013), no report of attrition rates (e.g. McBride et al., 2002) or had very high attrition rates (e.g. Tiwari, Avery, & Lai, 2003). Other studies did not control for confounders, such as participants' background (e.g. Dennett, 2014; Lee et al., 2011; Lun et al., 2010; Park, Niu, Cheng, & Allen, 2021; Yeh & Chen, 2003; Zhang & Zhang, 2013).

7.3 How do Chinese students' critical thinking skills compare with those of other nationalities?

Of the 15 studies, eight compared CT skills of Chinese students with those of other nationalities, and six examined CT dispositions. Only one study focused on CT styles.

CT skills include interpretation, analysis, evaluation, inference, explanation, deduction, and assumption. Some studies evaluated only a subset of these skills, while others included all these skills, depending on the test instruments used. Among the eight studies reviewed, four showed mixed results, three indicated that Chinese students exhibited higher CT skills than American students, and one suggested Chinese students

possessed lower levels of CT skills compared to New Zealand European students (see Table 7.2).

Table 7.2 Summary of results on literature comparing critical thinking skills (n = 8)

Higher CT skills	Lower CT skills	No difference	Mixed	Strength of evidence
				4*
			1	3*
1			1	2*
2	1		2	1*

Mixed results (n = 4)

Four studies reported mixed outcomes, indicating that Chinese students outperformed peers from certain nationalities in CT skills, but not others, and only in specific sub-skills.

Loyalka et al. (2021) compared the CT skills of Chinese, Indian, Russian, and American students in two disciplines (computer science and electrical engineering). Using the HEIghten® suite of assessments from Educational Testing Service, the study found that Chinese and American students had similar CT scores in the first two years, while Indian and Russian students had lower scores. In the fourth year, Chinese students performed similarly to Russian students but higher than Indian students. However, compared to American students, Chinese students performed worse. While American students improved their CT skills in the last two years, students from other countries showed a decline, with Chinese students showing the most significant decline.

This study, involving over 30,000 students across four countries, is the largest in this area. It is rated 3* on account of the large number of participants and the careful consideration given to the choice of instrument, languages, and testing environment. The measurement instrument was designed to be culturally neutral, and students were tested in their native language versions of the CT test.

Despite the care taken to ensure cultural comparability, several weaknesses lower the strength of the evidence to 3*. The number of participating students in each country

was highly unequal, with considerably more Indian ($n = 17,455$) and Chinese ($n = 9,247$) students than Russian ($n = 4,703$) and American ($n = 973$) students. Although sampling weights were adopted to address this imbalance, potential biases in sample selection remain. In India, Russia, and China, students were selected by random sampling, whereas in the US, students volunteered. Weighting for unequal sample sizes could amplify bias, particularly as American students were self-selected. Another issue is the high attrition rate among US students (39%). Although the researchers acknowledged the missing data and addressed it by including missing value dummies in the regression, such replacements cannot usually overcome the bias introduced. Missing cases and data are seldom random (Gorard, 2020). Those who drop out or do not answer certain questions are likely to be different from those who do. For example, it is possible that those who drop out, or did not complete the test may be weaker students. Using weighting to address missing cases among the US cohort may, in fact, magnify the bias. Additionally, the study was limited to only two disciplines (computer science and electrical engineering), which restricts the generalisability of the findings. Furthermore, there was a gender imbalance in the sample, with more than 60% of participants from China, India, and Russia being male. This gender difference could help explain the differences between groups, as gender is an important variable in measuring CT skills (Ennis et al., 2005b).

Hu et al. (2020), a 2* study, also showed mixed results. The study compared British and Chinese final-year accounting and finance students in a British university. While Chinese students scored marginally higher than British students in inferential skills (55% vs 51%), they performed much worse than British students in tests of assumption, arguments, and interpretation. On the test of deduction, Chinese students were on par with British students (62.5% vs 63%). The overall composite scores of Chinese students were lower than those of British students. This study was rated 2* because of the relatively small number of cases (50 in each group). It is also unclear how the students were selected. Moreover, a short version of the Watson-Glaser Critical Thinking Appraisal questionnaire (WGCTA) Form S was used, which was translated into Chinese. This may introduce a possibility of error in translation. The testing process was also problematic. Chinese students were initially tested using the English version of the test, and then the Chinese version. Both included the same content, likely leading

to familiarity with test items. Additionally, lecturers (who were not blinded), rather than researchers, administered the test, which may introduce potential problems including inconsistency of the research setting and unconscious bias (e.g. teachers may unconsciously give students greater support knowing that their scores will be compared). There is also the element of teacher expectation, which can affect student performance in the test.

Two other lower rated studies (1*) also suggest mixed results. **Dong et al.'s (2010)** study, for example, showed that although Chinese undergraduates had an overall higher score in CT skills (mean 19.20, SD 4.32) than students in US higher education (mean 16.80, SD 5.06), they performed lower in skills of analysis and induction. This study was rated 1* because of the very small number of Chinese cases ($n = 25$), the majority of which were males ($n = 17$).

Liu (2013) employed a similar research design and measurement instrument as Dong et al. (2010) comparing Chinese university students with the norm of four-year college and university students in the US on the California Critical Thinking Skills Test (CCTST), but focused on second-year Chinese undergraduates ($n = 30$) from two English programme classes at one University in China. Chinese students' overall CT skills scores (mean 19.83) are higher than those of American students (mean 16.80). However, unlike Hu et al. (2020) who reported that Chinese students performed better than British students on inferential skills, Chinese students performed worse on inferential skills compared to American students. There are some differences between the two studies. In Hu et al.'s (2020) study, students were final-year accounting and finance students, while those in Liu's (2013) study were second-year English programme students, most of whom were science majors. These students are, therefore, not representative of the average Chinese university students. It is also unclear if these students were compared with the general American undergraduate population, and whether the American and Chinese students were similar in terms of age and other demographic characteristics. Therefore, the study was rated 1* because this finding can only suggest a small advantage for Chinese students, but the results are far from conclusive given the lack of a similar comparator.

Higher CT skills (n = 3)

Three studies assessed Chinese students as having better CT skills than American students. Only one was rated 2*, the other two were 1*.

Ku et al. (2006) recruited 142 Chinese students from a premier Hong Kong university and 153 American students from a public university in southern California. CT was assessed using the Halpern Critical Thinking Assessment Using Everyday Situations (HCTAES) and translated to Chinese for the Chinese students. Chinese students scored higher than US students on the overall CT score. However, the results of the individual sub-scales were not reported. Additionally, some background elements such as admission criteria and undergraduate major were not controlled for. It is uncertain whether the two universities have similar levels of admission standards. There may be pre-existing differences between the students in the two institutions which were not accounted for. For example, around 77% of American students majored in social science, whereas only 40% of the Chinese students did. The disparity in majors is likely to influence CT skills performance (Bailin et al., 1999). Therefore, this study was rated as 2*.

Zhang and Zhang (2013) draw a similar conclusion. This study was rated 1* because it used Pintrich et al.'s (1991) Motivated Strategies for Learning Questionnaire (MSLQ) to measure CT, which is unusual. MSLQ is designed to measure motivation and learning strategies rather than CT skills. In this study, 197 Chinese university students from an English class and 165 American students from communication classes completed the test. To mitigate the influence of language, the test was translated into Chinese (with an alpha reliability of 0.90). Their results suggested that Chinese students performed better (mean 3.67, SD 0.92) than American students (mean 3.24, SD 0.87). However, the report did not explain the sample selection process or account for non-responses. Comparing students from English classes with those from communication classes is questionable, as these groups may differ in terms of entry qualifications. More demographic information would have been helpful. Using an instrument developed for Western education systems to measure a completely different construct casts doubt on the internal validity of the assessment. Another potential threat to the study's credibility

is that the American cohort received extra credit for participation, whereas the Chinese students did not receive similar incentives. Hence, this study was rated 1*.

Another 1* study (**Park et al., 2021**) reported that Chinese university students display higher CT skills than their American counterparts. The aim of this study was to investigate influence of culture on the CT skills of Chinese ($n = 166$) and American students ($n = 103$). CT was assessed using a combination of two vignettes from Lawson, Jordan-Fleming, and Bodle's (2015) Psychological Critical Thinking Exam, five questions from CCTST, and one vignette from the Sternberg Scientific Inquiry and Reasoning. Both open-ended and close-ended questions were included in their CT test. The final scores were averaged from these three assessments. The study found that Chinese students scored higher in CT skills (mean 1.32, SD 0.59) than their American peers (mean 1.02, SD 0.44).

This study was rated 1* because the two groups compared were not equivalent. Some students had advanced research experience, while others had none, but it is not clear what proportion in each group had advanced research experience. Research experience is positively correlated with CT skills (Haritania et al., 2019), indicating that those with more research experience may initially have better CT outcomes. Since the samples were not randomised, the proportion of students with and without research experience in each group may differ, which could partially explain the difference in CT performance. Furthermore, their test only considered several key dimensions of CT, including evaluation, logical reasoning, and probabilistic thinking, omitting other aspects such as analysis and deduction. Using the average scores of the three tests may not accurately measure general CT skills. For example, the test on scientific reasoning and inquiry may favour those with extensive research experience. It would be more informative to consider the weight of each test item.

Lower CT skills ($n = 1$)

Only one study reported a negative result.

Lun et al. (2010) compared the CT skills of 24 Chinese university students with 35 New Zealand students of European descent. The study included Asian students, with

Chinese students as a subset. The closed-ended section of the Halpern Critical Thinking Assessment Using Everyday Situations (HCTAES) was used as a measure of CT skills. Chinese students performed worse (mean -1.26, SD 1.70) than New Zealand European students (mean 0.87, SD 1.13). However, the small number of non-representative participants from one university in New Zealand means the results cannot be generalised to the wider Chinese student population, especially since the Chinese students were recruited from an international university rather than local Chinese universities. Another issue is that all participants were tested in English. The extent to which language may have impeded the performance of Chinese students, whose first language is not English, is unknown. Additionally, the students were asked to self-report their English proficiency, which is not a reliable measure. Therefore, the study was weak in evidence and rated 1*.

Overall, there is no conclusive evidence about Chinese students' CT skills either way. The stronger studies suggest mixed results.

7.4 How do Chinese students' critical thinking dispositions compare with those of other nationalities?

CT disposition refers to an internal tendency influencing one's beliefs or actions (Facione et al. (1995). Essential elements of CT disposition include truth-seeking, open-mindedness, and inquisitiveness (Ennis, 1985; Facione, 1990; Halpern, 1998).

The review identified six studies comparing the CT dispositions of Chinese students with those of other nationals. Four studies indicated that Chinese students had lower CT dispositions (Lee et al., 2011; McBride et al., 2002; Tiwari et al., 2003; Yeh & Chen, 2003), one showed no difference (Dennett, 2014), and one presented mixed results (Petrini & Kawashima, 2003). No studies indicated that Chinese students had higher CT dispositions. Although most studies suggest lower CT dispositions among Chinese students, this finding is not substantiated by the stronger studies. The evidence is therefore inconclusive.

Table 7.3 Summary of results on literature comparing critical thinking dispositions (n = 6)

Higher disposition	Lower disposition	No difference	Mixed	Strength of evidence
				4*
				3*
				2*
	4	1	1	1*

Lower CT dispositions (n = 4)

Four studies indicated that Chinese students have lower CT dispositions compared to other nationals, but their evidence is weak. All were rated 1*.

Lee et al. (2011) compared the CT dispositions of Chinese nursing students (n = 407) and Korean nursing students (n = 355), although they presented their study as an evaluation of CT skills. Chinese students demonstrated marginally lower levels of CT dispositions (mean 94.43, SD 7.26) than Korean students (mean 95.60, SD 8.59). The study was conducted in two Korean universities (four-year) and two Chinese universities (five-year).

Although this study attempted to track changes in students' CT dispositions, it did not examine the same cohorts across years. Instead, it compared first-year students with final-year students and found that gains in CT scores between first-year and final-year students were greater for Korean students than for Chinese students. They concluded that Korean students made more significant improvements over time. However, this conclusion was problematic. First, this was not a longitudinal study, and any difference between first-year and final-year students could simply be a reflection of the quality of students between cohorts. Second, the final-year Chinese students at the participating universities had one more year of university education compared to their Korean counterparts due to differences in the duration of university programs. The comparison was not equivalent. Third, as acknowledged by the authors, the Korean freshmen were already exposed to a CT course, whereas the Chinese students received no CT-related curriculum. Therefore, the study was rated 1*.

McBride et al. (2002) compared the CT dispositions of Chinese (n = 234) and American (n = 218) physical education students using the California Critical Thinking Dispositions Inventory (CCTDI). They reported that Chinese undergraduates scored lower in truth-seeking, inquisitiveness, maturity, and self-confidence. However, scores for analyticity, systematicity, and open-mindedness were not reported due to low Cronbach's alpha coefficients for the Chinese samples. The authors should have investigated why these constructs had low reliability instead of ignoring them. Although the cohort sizes were comparable, Chinese students were drawn from one university, while American students came from nine institutions. Any difference in CT disposition could be due to the kind of students in the one Chinese university, which is not representative of Chinese students in higher education in general. Additionally, inconsistencies in data reporting, such as the mean score for maturity being listed as 39.35 in the table but 30.35 in the text, weaken the study's credibility. Consequently, this research was rated 1*.

Tiwari et al. (2003) focused on nursing students, comparing the CT dispositions of Hong Kong Chinese students (n = 222) and Australian students (n = 162). Their results indicated that Chinese students had lower CT dispositions (mean 268.36, SD 21.58) than their Australian counterparts (mean 287.73, SD 30.98). Chinese nursing students scored lower in all seven sub-sets of CCTDI: truth-seeking, open-mindedness, analyticity, systematicity, self-confidence, inquisitiveness, and maturity. The study was rated 1* due to the lack of control for age. Although the authors claimed that both Chinese and Australian students were similar in age, the age of the Australians was not reported. Hence, it remains unknown whether the differences in CT dispositions are attributable to age or cultural differences. Another factor affecting the robustness of the results is the low response rate (61% for Chinese and 49% for Australian students), potentially introducing non-response bias (Sheikh & Mattingly, 1981).

Yeh and Chen (2003) found that Taiwanese nursing students (n = 214) scored lower than American nursing students (n = 196) on six sub-scales of the CCTDI: truth-seeking, open-mindedness, analyticity, systematicity, self-confidence, and maturity, except for inquisitiveness. However, this study was rated 1* because there were differences between groups that were not accounted for. For example, the Taiwanese students were

younger (mean age 22) than their American counterpart (mean age 28). Previous studies have shown that CT dispositions correlate with age (Emir, 2009). Therefore, the difference in CT dispositions between the two groups may be due to age rather than nationality. Additionally, almost half of the American students in this study had previous nursing experience (45.6%), whereas only 7.7% of Taiwanese students did. Nursing experience has also been found to be positively correlated with CT dispositions (Feng et al., 2010). This experience disparity may explain the lower CT disposition scores of the Taiwanese students. Another issue, as with other studies on CT dispositions, is the self-report nature of the CCTDI. While the tool is a standardised and independent measurement of CT dispositions, self-report is notoriously unreliable (Paulhus & Vazire, 2007; Slavin, 2017). The convenience sampling further reduced the reliability of this comparison due to potential self-selection bias.

No difference (n = 1)

Contrary to the above studies, **Dennett (2014)** found no difference in CT dispositions between Chinese and American students. However, the evidence is weak (rated 1*) due to the small, non-random sample of students from one university. Only 41 Chinese and 50 American students participated, all from the same American university. Moreover, the English version of the CCTDI was used for both groups. Using an English CT assessment for Chinese students is inappropriate, as language competency could influence performance (Floyd, 2011; Hu et al., 2020; Manalo & Sheppard, 2016). Any difference in CT performance could be attributed to the different level of language proficiency of the two groups. It is also problematic to compare Chinese students studying in America with American home students as Chinese international students who have chosen to study abroad may be a biased group. They are likely to be more open-minded, are probably higher-performing students from well-to-do families. They are therefore not representative of Chinese students in general.

Mixed results (n = 1)

Things are more complex when Chinese students' CT dispositions are compared with those of learners from more than one country. **Petrini and Kawashima (2003)** compared the CT dispositions of Chinese, Japanese, and Samoan nursing students though they claimed their study was an evaluation of CT skills. As with Lee et al.'s

(2011) study, this study also used an instrument designed to measure CT dispositions to assess CT skills. The study found that compared to Japanese students, Chinese students scored higher on analyticity, systematicity, and self-confidence, but lower in truth-seeking and open-mindedness. However, there was no difference in Chinese students' CT dispositions when compared to Samoan nursing students. This study was rated 1* due to some methodological weaknesses. Firstly, the use of convenience samples meant that those who took part were likely to be volunteers or self-selected individuals. Hence, the findings cannot be reliable as the groups compared are not equal. Secondly, the cases compared were highly unbalanced. There were 165 Japanese, 300 Chinese and 70 Samoan. It is also unclear how students were recruited. Thirdly, while all the students in the three countries were females, they differed in terms of age and work experience. For instance, the Chinese students ranged in age from 21 to 25 and all of them had little clinical experience. Samoan students, on the other hand, ranged in age from 16 to 62, with extensive nursing experience. The groups cannot be representative of the population studied. The failure to control these background elements casts doubts on the reliability of the findings.

In summary, the evidence on the level of CT dispositions of Chinese students is weak. Based on existing research, there is no evidence that Chinese students' CT dispositions are higher or lower than other nationals due to the small number of studies all of which are methodologically unsound — e.g. comparing samples of vastly different age, experience and background.

7.5 How do Chinese students' critical thinking styles compare with those of other nationalities?

CT style refers to the way an individual demonstrates or practises CT (Lamm, 2015). Two types of CT styles have been identified and assessed: information seeking and engagement (Lamm & Irani, 2011). Information seekers acknowledge their limitations in knowledge or experience and strive to gain more information before solving problems. Engagers exhibit a desire to communicate and show confidence in explaining their reasoning process when making decisions. According to Lamm and Irani (2011), a good critical thinker is one who embodies both styles.

Table 7.4 Summary of results of literature comparing critical thinking styles (n = 1)

Information seeking	Engagement	No difference	Mixed	Strength of evidence
				4*
				3*
1				2*
				1*

In this review, only one study (**Lu et al., 2021**) meeting the inclusion criteria considered students' CT styles. This study compared the CT styles of 104 US students (37 males) and 103 Chinese students (69 males) majoring in agriculture. CT styles were measured using the University of Florida Critical Thinking Inventory (UFCTI, see Chapter 2), which was translated into Chinese for the Chinese cohort (Cronbach alpha 0.92). Unlike instruments measuring CT dispositions and skills, the UFCTI evaluates students' preferences for expressing CT and their behaviours (Lamm & Irani, 2011). It is important to note that the UFCTI does not measure the level of each style. A low score does not indicate a lower or higher level of CT styles (Lamm, 2015). Rather, it is a measurement of students' preferences.

The study revealed that the mean overall score for Chinese students was 80.67 (SD 4.96), while for American students, it was 77.87 (SD 5.05). According to UFCTI guidelines, students with an overall score above 79 are classified as seekers, whereas those below 78 are classified as engagers. These results suggest that Chinese students prefer information seeking, whereas American students are more inclined towards an engaging CT style. While this does not indicate whether American or Chinese students possess higher levels of CT, the different styles may help explain why Chinese students, on average, score lower on the CT skills test, which measures analytical, evaluative, and deductive skills.

The study was assessed as 2* due to several limitations. It did not control for potential confounding factors. For example, most Chinese participants were male, whereas most US students were female. There is evidence suggesting that gender may influence CT styles (Ennis et al., 2005b), and this factor cannot be overlooked. Another issue is the measurement quality, as the UFCTI relies on self-reported data, which is not an

objective measure. Students' CT styles may be related to their CT dispositions and skills; however, this review found no studies attempting to link these measures. Future studies could explore the relationship between CT styles and CT skills.

7.6 Chapter summary

To establish the claim often made by Western academics that Chinese students are lacking in CT, the review evaluates studies that compare the CT of Chinese students with those of students of other nationalities. Therefore, this review only considers studies that include a comparison group.

Overall, there is no evidence to support the claim that Chinese students have higher or lower CT skills than students from other countries. Research in this area is limited and of poor quality. While Chinese students may seem to be less disposed to CT, none of studies in CT dispositions has strong evidence. Only one study examined CT styles, indicating that Chinese students are more inclined to information seeking than their peers in the US.

Most studies reviewed had methodological weaknesses, ranging from small sample sizes, high attrition or low response rates, the use of convenience sampling to poor analytical processes. Most studies did not account for confounding variables or establish group equivalence. All these issues meant that the findings are not reliable, and we need to be cautious about the interpretations of the results.

The overall conclusion is that we still do not have a definitive answer. There is no conclusive evidence that Chinese students have higher or lower CT skills or dispositions compared to their peers from other countries. There are insufficient high-quality studies to draw any valid conclusions. These findings suggest the need for more robust, large-scale experimental studies. Research in this field needs to improve.

Section IV Results of the primary research

This section consists of four chapters. It presents the results of the primary research, which aims to evaluate the effectiveness of EnglishFusion on Chinese secondary students' critical thinking (CT) skills and academic attainment. It addresses the following research questions:

1. Can critical thinking skills be taught to Chinese secondary students who are not traditionally exposed to critical thinking?
- 2a. Does EnglishFusion improve Chinese secondary students' critical thinking skills?
- 2b. Does EnglishFusion have a differential impact on the critical thinking skills of sub-groups of students (by age, birth sex, ethnicity, prior academic attainment, prior critical thinking skills, schools, parental involvement in children's education, and home background)?
- 3a. Does EnglishFusion improve Chinese secondary students' academic performance?
- 3b. Does EnglishFusion have a differential impact on the academic attainment of sub-groups of students (by age, birth sex, ethnicity, prior academic attainment, prior critical thinking skills, schools, parental involvement in children's education, and home background)?
4. Does training and teaching EnglishFusion alter teachers' critical awareness and attitudes towards teaching critical thinking?

Chapter 8 presents the research findings of the pilot study, including both the impact evaluation on the intervention's effectiveness and the process evaluation on how the intervention and the study could be improved. The subsequent three chapters provide the results of the main trial. Chapter 9 focuses on the impact evaluation on CT skills, while Chapter 10 presents the impact evaluation on academic achievement. Chapter 11 reports the process evaluation, which supplements the impact evaluation results and explores how the intervention is actually implemented in Chinese secondary English language classrooms.

Chapter 8 Results of the pilot study

Prior to the main study, a pilot study was conducted to assess the feasibility of the research design, randomisation process, test instruments, questionnaire design, and intervention design (including lesson plans and activities). It also evaluated the training of teachers and allowed for rehearsal of the analysis. Feedback from both students and teachers was used to inform the development of the main trial. The results of the pilot study are detailed in this Chapter.

8.1 Revisions for the main study

EnglishFusion (the name of the intervention) was delivered by one experimental teacher with seven years of experience teaching English in secondary schools. All six CT lessons were integrated into the usual English course, with students and the teacher in the same classroom, engaging in face-to-face communication. Although it was initially planned for one CT lesson per week over six weeks, school closures due to COVID-19 disrupted the schedule, and the intervention was ultimately extended. Additionally, the last three CT lessons were noticeably shorter due to the lockdown, which resulted in a reduced regular school schedule.

Following the implementation of the pilot study and considering feedback from students and teachers, revisions were undertaken to enhance the quality of the main trial. A summary of key differences is reported in Table 8.1.

Table 8.1 Summary of main differences between the pilot study and the main trial

	The pilot study	The main trial
Participants	122 students, two English language teachers, one school	2,055 students, 21 English language teachers, four schools
Teacher training	1. Self-learning and informal communication via telephone 2. Only presented with the lesson plans and slides	1. Formal training sessions and informal weekly followed-up training, face-to-face 2. One lesson demonstration added 3. Discussions with other experimental teachers

Duration and frequency	<ol style="list-style-type: none"> 1. One month (during Covid-19 pandemic) 2. Irregularly, the last three lessons were merged 3. Largely dependent on the teacher's availability 	<ol style="list-style-type: none"> 1. Three months 2. Regular implementation, once a week 3. All experimental teachers followed the same teaching pace
Content of intervention	Six CT lessons only	<ol style="list-style-type: none"> 1. Seven CT lessons 2. Six small CT tasks closely integrated with textbook materials
Delivery of the intervention	<ol style="list-style-type: none"> 1. English only in slides 2. Few pictures 3. No handout for students 	<ol style="list-style-type: none"> 1. Added Chinese translation 2. More pictures 3. Handouts for each student
Fidelity assessment	<ol style="list-style-type: none"> 1. Class observations via audio-recordings 2. Remote interviews with students and the teacher 	<ol style="list-style-type: none"> 1. On-site class observations 2. Face-to-face informal communication and interviews with students and teachers
Measurement	<ol style="list-style-type: none"> 1. Modified CT tests 2. Teachers delivered the pre and post CT tests 3. The latest term's final examination score collected to examine students' academic background 	<ol style="list-style-type: none"> 1. Modified CT tests with balanced difficulty 2. The researcher administrated the pre and post CT tests; teachers were unaware of the testing content 3. Final examination scores collected from two terms and administered by the local education department
Outcomes	<ol style="list-style-type: none"> 1. Primary outcome: CT skills 2. No secondary outcomes 	<ol style="list-style-type: none"> 1. Primary outcome: CT skills 2. Secondary outcome: academic attainment including Chinese, Maths and English scores

8.2 Impact evaluation

Although the main aim of the pilot study was not about the impact of EnglishFusion, a brief discussion of the results is necessary to evaluate whether the intervention was effective in students' CT skills. As this was a pilot study, risks and biases could be associated with the sample size (N = 122).

The experimental group had a lower level of CT skills compared to the control group at pre-intervention (see Table 8.2). Compared to the control group, they performed worse at identifying assumptions (ES = -0.43) and interpreting messages (ES = -0.23). Interestingly, they performed better at making justified arguments than the control group (ES = +0.30). There was little difference between the groups in terms of making deductions and understanding inferences.

Table 8.2 Comparison of pre-test critical thinking scores between experimental and control groups (N = 122)

	Experimental group (n = 61)		Control group (n = 61)		Overall (N = 122)		ES
	Mean	SD	Mean	SD	Mean	SD	
Argument	2.34	0.79	2.08	0.90	2.21	0.86	0.30
Assumption	1.95	0.83	2.30	0.78	2.12	0.82	-0.43
Deduction	2.57	0.69	2.59	0.64	2.58	0.67	-0.03
Inference	1.46	1.03	1.41	0.99	1.43	1.00	0.05
Interpretation	1.15	0.68	1.31	0.72	1.23	0.70	-0.23
Overall CT skills	9.48	1.88	9.69	1.78	9.58	1.82	-0.12

In the post-test, the experimental group also showed lower CT skills. They performed worse than the control group in tests of deduction, inference and assumption, but outperformed them slightly in argument and interpretation (see Table 8.3).

Table 8.3 Comparison of post-test critical thinking scores between experimental and control groups (N = 122)

	Experimental group (n = 61)		Control group (n = 61)		Overall (N = 122)		ES
	Mean	SD	Mean	SD	Mean	SD	
Argument	1.44	0.99	1.38	0.86	1.41	0.92	0.07
Assumption	2.03	0.80	2.10	0.77	2.07	0.78	-0.09
Deduction	1.46	0.87	1.75	0.91	1.61	0.90	-0.32
Inference	1.31	0.99	1.54	0.85	1.43	0.93	-0.25
Interpretation	1.61	0.78	1.56	0.83	1.58	0.80	0.06
Overall CT skills	7.85	1.90	8.33	2.06	8.09	1.98	-0.24

Since the two groups were clearly not balanced at the outset, any difference in CT skills after the intervention could be attributed to the unbalanced baseline. To ensure that the difference in CT skills post-intervention is not due to pre-existing differences, the impact evaluation used the students' gain scores.

The gain scores reveal a mixed outcome (see Table 8.4). Compared to the control class, the experimental class showed greater progress in identifying assumptions and interpreting messages. However, they demonstrated negative progress in making deductions, inferences and arguments.

Table 8.4 Comparison of gain in critical thinking scores between experimental and control classes (N = 122)

	Experimental group (n = 61)		Control group (n = 61)		Overall (N = 122)		ES
	Mean	SD	Mean	SD	Mean	SD	
Argument	-0.90	1.33	-0.70	1.27	-0.80	1.30	-0.15
Assumption	0.08	1.07	-0.20	1.09	-0.06	1.09	0.26
Deduction	-1.11	1.05	-0.84	1.10	-0.98	1.08	-0.25
Inference	-0.15	1.40	0.13	1.27	-0.01	1.34	-0.21
Interpretation	0.46	0.96	0.25	1.03	0.35	1.00	0.21
Overall CT skills	-1.62	2.32	-1.36	2.83	-1.49	2.58	-0.10

EnglishFusion does not appear to have a promising effect on improving students' CT skills, but this should be interpreted with caution. First, there were only two teachers (one in each group), and any differences in outcomes between groups could be attributed to teacher characteristics. The experimental teacher was older than the control teacher, aged 32 and 27 respectively. The control teacher was a recent graduate with one year of teaching experience, while the experimental teacher had accumulated seven years of teaching experience. The older, more experienced teacher was observed to adhere more closely to traditional teaching methods, whereas the younger teacher showed greater openness to new ideas and pedagogical approaches.

Second, a low fidelity of the intervention was observed in the experimental class. The experimental teacher still adhered to traditional, teacher-centred teaching methods. She explained too much and provided limited time for student discussion and idea sharing. The insufficient preparation of the intervention delivery also posed a problem. Students often expressed doubts when there were incorrect translations and unclear explanations. Additionally, the planned sections, including summary and homework, were not completed in class.

Third, some CT components had already been somewhat integrated into the control classes. To address the question of whether friends should be the same or different, the control teacher not only asked students to identify the textbook authors' opinions but also required them to provide supporting details to justify answers. Although the usual English lesson did not extend beyond the formal textbook, it was reasonable as students were expected to achieve language skills through textbook learning. What mattered was how the teacher used these course materials to facilitate her students' thinking.

8.3 Chapter summary

The pilot study could inform necessary modifications for the main trial. Based on feedback on the delivery of the intervention, teaching materials, test instruments, and questionnaires, corresponding revisions were made to improve the robustness of the main trial that evaluates the effectiveness of EnglishFusion.

The pilot study also served as a rehearsal for the analytical procedures. The findings suggest that the infusion CT method was not promising in promoting the CT skills of Chinese students. However, due to the initial imbalance in CT skills between the experimental and control groups, this result cannot be conclusive. Additionally, the negative impact evaluation on CT skills could be attributed to teacher differences, low fidelity of the intervention, and diffusion issues.

In the main trial, corresponding changes were made to improve the rigour of the study. The teacher difference was addressed by recruiting more than two teachers in the main trial. Follow-up teacher training and sufficient teaching resources were provided to increase intervention fidelity. Experimental teachers and students were reminded to keep the intervention materials confidential to prevent any diffusion problems.

Chapter 9 Impact of EnglishFusion on critical thinking

The primary research evaluates the effectiveness of EnglishFusion on students' critical thinking (CT) skills and academic attainment. This chapter focuses on the impact of EnglishFusion on CT skills. Chapter 10 discusses the results of the impact on academic outcomes. As a reminder, the research questions for the impact evaluation are:

1. Can critical thinking skills be taught to Chinese secondary students who are not traditionally exposed to critical thinking?
- 2a. Does EnglishFusion improve Chinese secondary students' critical thinking skills?
- 2b. Does EnglishFusion have a differential impact on the critical thinking skills of sub-groups of students (by age, birth sex, ethnicity, prior academic attainment, prior critical thinking skills, schools, parental involvement in children's education, and home background)?
3. Does training and teaching EnglishFusion alter teachers' critical awareness and attitudes towards teaching critical thinking?

9.1 The sample

The trial involved 2,055 Grade eight (aged 13-14) students from four secondary schools in Sichuan Province, China. The four secondary schools were purposefully selected by the local Department of Education due to their differences in school size and the education levels they covered. These schools were located in a rural area, making them somewhat disadvantaged compared to those in larger cities. They generally had poorer teaching facilities and resources. They were all in the same county. As different English textbooks are used in different regions of China, students from the same area were taught the same curriculum content and sat for the same final examinations.

Regarding school size, School A was particularly large, with 22 classes (1,152 students) in Grade eight, while School D was the smallest, with only three classes (108 students). These schools covered different educational levels. For instance, both secondary and high school levels were included in School A, and the same applied to School B. However, School C taught students from preschool, primary, and secondary levels, whereas School D only covered the secondary education level. Overall, the sample

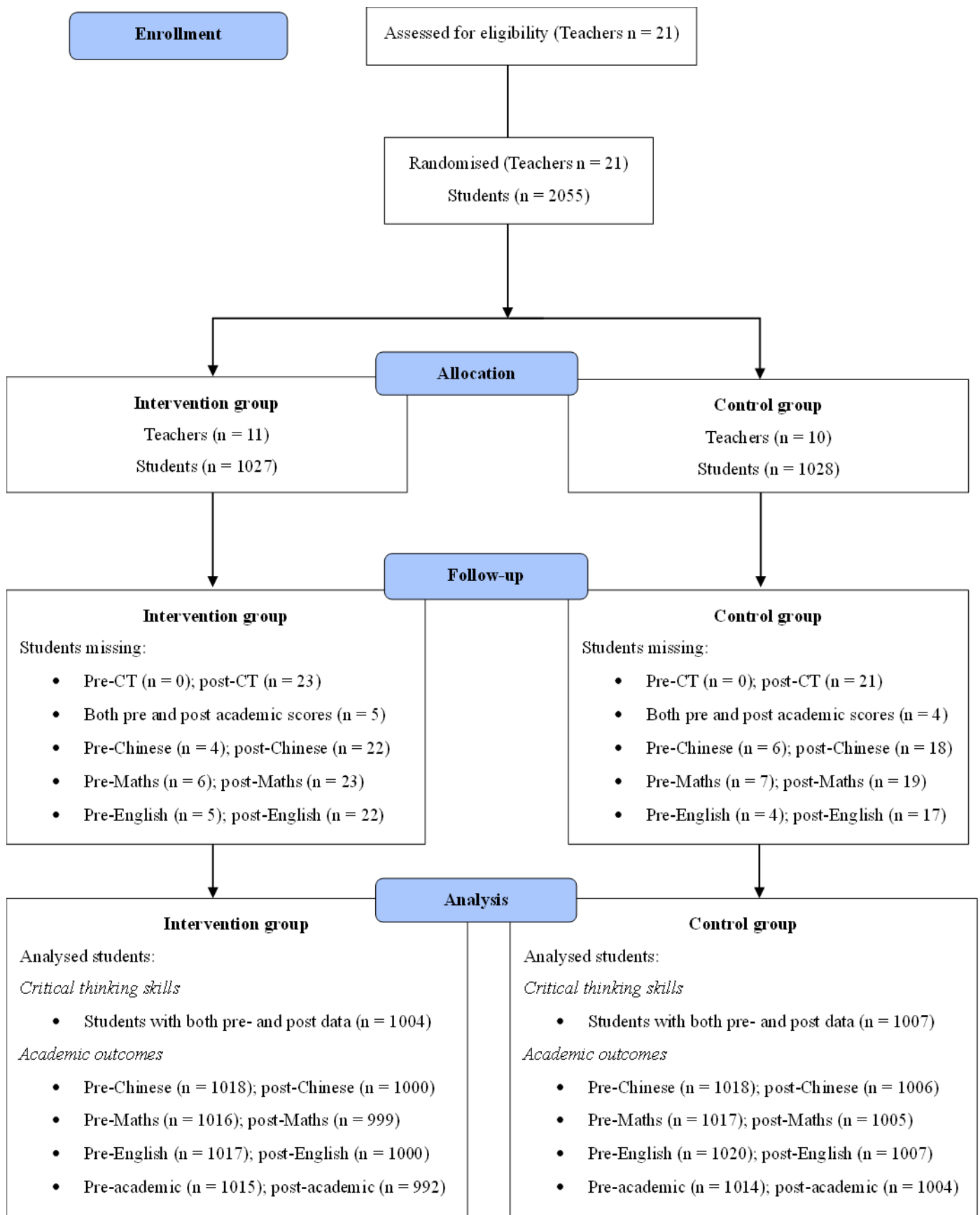
included 1,152 students and 11 teachers from School A, 638 students and 6 teachers from School B, 157 students and 2 teachers from School C, and 108 students and 2 teachers from School D (see Table 9.1).

Table 9.1 Characteristics of the four participating schools

Schools	School size	Education levels covered	Number of Grade 8 classes	Number of Grade 8 teachers	Number of Grade 8 students
A	Large	Secondary and high school	22	11	1,152
B	Large	Secondary and high school	10	6	638
C	Medium	Preschool, primary and secondary	3	2	157
D	Small	Secondary level only	3	2	108
Total	\	\	38	21	2,055

The CONSORT flow diagram (see Figure 9.1) describes the numbers of teachers and students from enrolment to analysis. Initially, these four schools were contacted, and all agreed to participate in the study. All English language teachers and their Grade 8 students in these schools took part. Teachers within the school were then randomly assigned to either the experimental group which will receive EnglishFusion lessons or to the control group which will continue with the regular curriculum (see Chapter 6). This results in 11 teachers with 1,027 students being allocated to the experimental group, and 10 teachers with 1,028 students in the control condition.

Figure 9.1 The CONSORT flow diagram



Attrition

Prior to randomisation, all students completed the CT pre-test and the student questionnaire. However, during the three-month intervention period, 44 students dropped out. Specifically, in the experimental group, six students took long-term leave due to sickness, mental health issues, or other personal reasons, and 17 students left school due to transfers, suspensions, or dropouts. A total of 23 experimental students did not finish the CT post-test. In the control group, six students did not complete the post-intervention test because they were absent when it was administered, despite efforts to get them to complete it later. Another 15 left school due to transfers, suspensions, or dropouts. Altogether 21 control students did not complete the CT post-test. The attrition in both the experimental and control groups was relatively balanced and unlikely to be attributable to the intervention itself.

In the final analysis, 1,004 experimental students from the experimental group and 1,007 control students were included in the impact evaluation on CT skills.

For academic attainment, nine students were missing both pre- and post-tests (see Figure 9.1), five were from the experimental group and four from the control. Academic attainment was assessed through two final examinations administered by the local department of education. Part of the attrition occurred because some students did not sit the examinations or were caught cheating during the tests. The examination environment was stringent, and students arriving late to the testing classroom were rejected from the test. Moreover, if students were found cheating, their scores were recorded as zero. As the exams are high-stakes on-time tests, students were not allowed a re-sit.

Additionally, ten students (4 experimental and 6 control students) were missing the pre-test for Chinese, 13 were missing the pre-test for Maths, and nine were missing the pre-test for English. At the end of the term, scores for 40 students' post-Chinese, 42 post-Maths, and 39 post-English were missing. It is clear that the number of missing data in post academic scores was almost four times larger than that in the pre academic grades. Moreover, the number of missing students for pre-academic scores in the experimental

group was similar to that of missing students in the control group, but there were more missing students in the experimental group for post-academic scores.

9.2 Characteristics of students

Age and birth sex

Students in both the experimental and control groups share similar age distributions, with the experimental cohort (mean age 14.01) slightly older than the control group (mean age 13.99).

The two groups also exhibit a relatively balanced distribution in terms of birth sex (see Table 9.2). There are slightly more girls than boys in the experimental group (521 vs 506 respectively), whereas in the control group, there are marginally more boys than girls (520 vs 508).

Table 9.2 Percentage of birth sex in experimental and control groups (N = 2,055)

	Experimental group (n = 1,027)	Control group (n = 1,028)
Male	49.3	50.6
Female	50.7	49.4

Ethnicity

The distribution of students from minority and majority ethnicities is also balanced between the two groups (see Table 9.3). Approximately 98% of all participants (n = 2,014) are from Han/majority ethnicity, with only 2% from minority backgrounds (n = 41). This is similar to the ethnic distribution in China. The experimental group has marginally more students from Han ethnicity (n = 1008) than the control group (n = 1,006), while the number of experimental students from minority backgrounds (n = 19) is slightly lower than that of control students (n = 22).

Table 9.3 Percentage of ethnicity in experimental and control groups (N = 2,055)

	Experimental group (n = 1,027)	Control group (n = 1,028)
Han	98.1	97.9
Minority	1.9	2.1

Home background

Family socioeconomic status (SES) is correlated to students' CT performance (Deal & Pittman, 2009; McCutcheon et al., 1992) and academic performance (Gorard & See, 2013; Liu & Lu, 2008). As data collection on pupils' household income and parental occupation was not permitted in China, this study used household possessions, as reported by the pupils, as proxy measures of SES. This method is used in international surveys such as PISA.

Table 9.4 Comparison of household possessions in experimental and control groups (N = 2,055)

	Experimental Group (n = 1,027)	Control Group (n = 1,028)	% differences
Own rooms	94.4	93.9	0.5
Study desks	89.6	89.7	-0.1
Computers for homework	23.6	24.2	-0.6
Wi-Fi	90.6	89.1	1.5
Bookshelves	67.5	68.1	-0.6
Classic literature	78.8	82.3	-3.5
Books of poetry	53.3	59.8	-6.5
Works of art (e.g. paintings)	35.2	38.1	-2.9
Books on art, music or design	30.5	34.5	-4
Musical instruments (e.g. pianos or guitars)	25.3	28.8	-3.5

Most of the students report having their own room, a study desk and Wi-Fi at home. There are notable differences between the groups in terms of cultural possessions. The control students appear to be from more cultured households, suggesting a higher level of cultural and intellectual engagement and appreciation of fine arts. They are more likely than the experimental students to possess books of poetry, books on art, music or design, classical literature, musical instruments and works of art.

Parental involvement in children's education

It is widely believed that parental involvement in children's education has an influence on their children's educational outcomes (Đurišić & Bunijevac, 2017), and perhaps their CT as well (Spence, 2012). Students were asked to indicate the frequency of their parent's involvement activities over the last academic year, rating from Never (0) to All the time (10).

In terms of parental involvement in their children's education, there is little difference between the two groups (see Table 9.5).

Table 9.5 Parental involvement in experimental and control groups (N = 2,055)

	Experimental group (n = 1,027)		Control group (n = 1,028)		Overall (N = 2,055)		Effect size (ES)
	Mean	SD	Mean	SD	Mean	SD	
School performance	6.10	3.22	6.15	3.12	6.12	3.17	-0.02
Homework	3.10	3.06	3.11	3.07	3.10	3.07	0.00
Political or social issues	4.48	3.58	4.22	3.47	4.35	3.53	0.07
Library or bookstore	2.83	3.16	2.91	3.11	2.87	3.13	-0.03
Reading	3.66	3.58	3.37	3.37	3.51	3.48	0.08
Overall	20.17	11.55	19.76	11.39	19.97	11.47	0.04

Experimental parents are slightly more likely to discuss political or social issues (ES = 0.07) with their children and talk about something they read (ES = 0.08) more often than control parents. Of all the parental involvement activities, children reported that their parents were more likely to discuss their school performance and political and social activities than their homework or reading.

Overall, the experimental and control groups are somewhat balanced with respect to age, birth sex, ethnicity and parental engagement in children's education. However, it is worth noting that the experimental group had a lower SES and poorer academic scores compared to the control group. To ensure a fair evaluation of the infused CT

lessons, it is important to control for these factors and use progress scores in both academic achievement and CT skills assessments.

9.3 Characteristics of teachers

Data on teachers' characteristics, including their age, birth sex, teaching experience and educational background was collected to see if they determine students' performance in CT and control for teacher differences in the regression analysis.

Age and birth sex

Teachers in the two groups were similar in age, with the control group slightly older than the experimental group (see Table 9.6). A greater age disparity was shown in the control group. The majority of teachers are female, with only two males, evenly distributed across the experimental and control groups.

Table 9.6 Comparison of teachers' age between experimental and control groups (N = 21)

	Experimental group (n = 11)	Control group (n = 10)
Mean	37.45	37.70
SD	7.43	10.52

Teaching experience

English teaching experience was measured by the number of years of teaching English. Experimental teachers have a shorter length of English teaching years than control teachers (see Table 9.7). Similar to the age distribution, the difference in years of English teaching experience is more noticeable within the control group.

Table 9.7 Comparison of English teaching experience between experimental and control groups (N = 21)

	Experimental group (n = 11)	Control group (n = 10)
Mean	15.50	16.50
SD	7.93	12.01

Educational background

Teachers in both groups are also similar in terms of educational background, with all except one graduating from normal universities (i.e. teacher training institutions). The majority (86%) hold undergraduate qualifications. Three teachers in the experimental group had Master's degrees. None of the teachers have studied overseas, indicating limited exposure to CT. Finally, despite the opportunity for additional comments or experiences, no further input was provided by the teachers.

9.4 Does EnglishFusion improve Chinese secondary students' critical thinking skills?

To recapitulate, the impact of EnglishFusion was measured by the standardised differences in the gain scores in the CT skills test between the two groups expressed as Hedge's *g* effect size. The primary outcome was students' CT skills, measured across five components: making arguments, identifying assumptions, making deductions, interpreting messages, and understanding inferences using the modified standardised CT tests that included 15 multiple-choice questions, with three items for each subset of CT. The scores for each section ranged from 0 to 3.

Primary outcomes

In the pre-test, the experimental group performed slightly better overall, and marginally worse on skills of argumentation, but ahead on assumption and interpretation skills (see Table 9.8).

Table 9.8 Comparison of pre-test critical thinking scores between experimental and control groups (N = 2,011)

	Experimental group (n = 1,004)		Control group (n = 1,007)		Overall (N = 2,011)		ES
	Mean	SD	Mean	SD	Mean	SD	
Argument	1.48	0.89	1.55	0.86	1.51	0.88	-0.08
Assumption	2.15	0.79	2.08	0.86	2.12	0.83	0.08
Deduction	1.53	0.81	1.50	0.91	1.52	0.86	0.03
Inference	1.23	0.88	1.21	0.90	1.22	0.89	0.02
Interpretation	1.38	0.78	1.30	0.82	1.34	0.80	0.10

Overall CT skills	7.76	1.88	7.64	2.20	7.70	2.05	0.06
-------------------	------	------	------	------	------	------	------

In the post-test, the experimental group outperformed the control group on three of the five skills (see Table 9.9). On the test of arguments, they were similar, but now the experimental group did slightly worse than the control on tests of interpretation.

Table 9.9 Comparison of post-test critical thinking scores between experimental and control groups (N = 2,011)

	Experimental group (n = 1,004)		Control group (n = 1,007)		Overall (N = 2,011)		ES
	Mean	SD	Mean	SD	Mean	SD	
Argument	1.65	0.82	1.64	0.86	1.64	0.84	0.01
Assumption	2.38	0.75	2.21	0.84	2.29	0.80	0.21
Deduction	2.04	0.84	1.88	0.86	1.96	0.86	0.19
Inference	1.48	0.92	1.30	0.84	1.39	0.88	0.20
Interpretation	1.31	0.86	1.34	0.90	1.32	0.88	-0.03
Overall CT skills	8.85	2.16	8.36	2.33	8.60	2.26	0.22

As the groups were not balanced at the pre-test, the gain scores between pre-and post-tests were used to estimate the impact of the intervention (see Table 9.10).

Table 9.10 Comparison of gain in critical thinking skills scores between experimental and control groups (N = 2,011)

	Experimental group (n = 1,004)		Control group (n = 1,007)		Overall (N = 2,011)		ES
	Mean	SD	Mean	SD	Mean	SD	
Argument	0.16	1.17	0.09	1.16	0.13	1.16	0.06
Assumption	0.24	1.00	0.12	1.11	0.18	1.06	0.11
Deduction	0.51	1.12	0.37	1.21	0.44	1.17	0.12
Inference	0.25	1.25	0.09	1.25	0.17	1.25	0.13
Interpretation	-0.07	1.13	0.04	1.18	-0.01	1.16	-0.09
Overall CT skills	1.09	2.59	0.71	2.90	0.91	2.76	0.14

The results showed that the experimental students made bigger gains in CT skills overall compared to the business-as-usual control group (ES = +0.14). The experimental group outperformed the control group on all sub-scales except for interpretation skills. Although the experimental group scored higher on the pre-test, they lagged behind on the post-test in this area. This suggests that EnglishFusion has not been effective in improving interpretation skills. It is possible that interpretation skills were not adequately addressed in the classroom. While other CT skills, such as recognising assumptions and drawing reasonable inferences, were emphasised and reinforced through the CT tasks, no specific lessons focused on interpretation. It is also possible that interpretation skills require more time to develop or that their effects may not become observable immediately. A long-term impact evaluation might be necessary to determine this.

Overall, there is a positive impact of EnglishFusion on students' CT skills, with the exception of the skill of interpretation.

Sensitivity analysis

The impact of the intervention has been measured using effect sizes, but an effect size does not take into account the scale of the study and any missing data. As there were students missing from both pre-and post-tests, the overall impact may be biased. To assess how secure the above findings are taking into account these missing cases, a sensitivity test, known as the Number Needed to Disturb (NNTD) was conducted (see Chapter 6). NNTD estimates the number of counterfactual scores (negative scores) that would be needed to make the effect size disappear. It is calculated as the “effect” size multiplied by the number of cases in the smallest group in the comparison. For this trial, the smaller group is the experimental group with 1004 students, and the effect size of CT skills' gain score is +0.14. Thus, the NNTD is 141 ($1004 * 0.14$). This calculation means that 141 cases would be needed to have negative results for the overall effect to be zero or become negative. This number is then compared with the number of missing cases. But, as the number of cases missing the CT skills test was 44, the findings are considered secure. In other words, if the missing cases were included and they all had negative effects, it would not be enough to alter the overall substantive results.

Although the missing cases are unlikely to alter the final findings, investigating their pre-test scores for CT skills provides further insight into the intervention’s impact. Table 9.11 shows that students in the experimental group with missing post-test scores had higher pre-test scores.

Table 9.11 Comparison of pre-test critical thinking scores for missing cases between experimental and control groups (N = 44)

	Experimental group (n = 23)		Control group (n = 21)		Overall (N = 44)		ES
	Mean	SD	Mean	SD	Mean	SD	
Argument	1.35	0.83	1.43	0.93	1.39	0.87	-0.09
Assumption	1.91	0.85	1.86	0.85	1.89	0.84	0.06
Deduction	1.57	0.84	1.29	0.85	1.43	0.85	0.33
Inference	1.35	0.88	0.95	0.86	1.16	0.89	0.45
Interpretation	1.35	0.71	1.19	0.87	1.27	0.79	0.20
Overall CT skills	7.52	1.68	6.71	2.19	7.14	1.96	0.41

While the differences in the scores of arguments and assumptions between the experimental and control groups are minimal, experimental students who missed the post-test performed better in other CT facets, including deduction, inference, and interpretation. Therefore, despite the balanced number of attrition in each group, the higher pre-test scores of experimental students who dropped out suggest that these students were more likely to be higher critical thinkers. Excluding them from the post-test could potentially pull down the overall impact.

To determine whether students who did not complete the post-test of CT differ from those who did, I compare the mean scores and standard deviations of overall CT skills. In other words, these missing cases’ pre-test scores of CT were examined against those of completed cases (see Table 9.12).

Table 9.12 Comparison of pre-test critical thinking scores between missing and completed cases

	N	Mean	SD
Experimental missing cases	23	7.52	1.68
Experimental completed cases	1004	7.76	1.88
Control missing cases	21	6.71	2.19
Control completed cases	1007	7.64	2.20

It appears that students who missed the post-test generally had lower pre-test scores than those who completed both tests. The mean pre-test score of control group students who missed the post-test was 6.71, lower than the 7.64 mean score of those who completed it (a difference of 0.93). Similarly, the mean pre-test score of experimental students who missed the post-test was 7.52 compared to 7.76 for those who completed it (a difference of 0.24). The difference between those missing and completed pre-test scores in the control group is much bigger compared to the experimental group. This suggests that if the 21 students in the control group missing the post-test were included, the overall CT score for the control group could have been lower, meaning the overall effect of EnglishFusion on CT skills could have been larger.

Additional analysis

To estimate the contribution of the intervention after accounting for students' background and prior CT score, a linear regression analysis was conducted using the post-test CT skills score as the dependent variable and background factors as predictors (or explanatory factors). As a reminder, these factors, including age, birth sex, ethnicity, household possessions, and parental involvement in education, were entered into the regression as the initial set of variables as these are factors beyond the students' immediate control. Students' prior CT score was included in the second block, followed by their prior academic attainment in the third block. Factors of teachers and schools were considered in the fourth block. Finally, student membership of the intervention group was entered to see how much receiving the intervention explains the variance in the outcome after controlling for students' background, prior CT level, prior academic performance, and school and teacher characteristics.

Table 9.13 Regression results predicting students' post-test critical thinking scores

Block	R	R Square	R square change
Classic literature, birth sex, ethnicity, discussing school performance	0.13	0.02	0.02
Pre-test CT score	0.21	0.04	0.02
Pre-test academic score	0.33	0.11	0.07
Teacher characteristics	0.33	0.11	0.00
Intervention	0.37	0.14	0.03

The results show that students' home background, as estimated by the possession of classic literature, birth sex, ethnicity and the frequency of parents discussing school performance, explains only 2% of the variance in post-test CT scores. While other background factors were considered, they were excluded from the regression model because of their minimal predictive power. The pre-test CT score explains an additional 2%, and the prior academic attainment adds 7% more variance. The teacher variable makes no difference. After controlling for these factors, the intervention explains an additional 3% variance in the post-CT scores.

Based on this model, the pre-test academic score is the strongest predictor of students' post-test CT scores. This is consistent with class observations and teachers' interviews where students with high academic achievement tend to be more interested and engaged in EnglishFusion lessons. They were eager to discuss and share ideas in class, which could help them increase their CT skills. The intervention also contributes to some small extent to students' CT skills.

While the R-squared value indicates the amount of variance in the dependent variable (i.e. CT post-test scores in this study) explained by the predictors (i.e. independent variables), the beta coefficient reflects the strength and direction of the relationship between each independent variable and the dependent variable, while holding all other variables constant. As the predictors (e.g. academic scores, prior CT scores and household items) have different units of measurement, the standardised beta coefficients are used to interpret the strength of association.

Table 9.14 Coefficients for the model predicting post-test critical thinking scores

	Unstandardised <i>B</i> coefficient	Standardised beta coefficient
Constant	5.03	-
Classic literature	0.05	0.01
Birth sex	-0.12	-0.03
Ethnicity	-0.83	-0.05
Discussing school performance	0.00	0.00
Pre-test CT score	0.12	0.11
Pre-test academic score	0.01	0.32
Teacher characteristics	0.00	0.00
Intervention or not	0.83	0.18

Table 9.14 shows that students' post-test CT score is most strongly related to their pre academic score (+0.32), followed by the intervention (+0.18) and pre-test CT score (+0.11). This means that for every standard unit increase in prior academic scores, students' post-test CT scores increased by 0.32 units accordingly. The intervention group did slightly better than the control (+0.18, where control is coded as 0 and intervention as 1).

Other background variables (classic literature, birth sex, ethnicity and discussing school performance) are weakly related to the post-test (less than 0.10). Teacher characteristics contribute little to student's CT skills.

9.5 Does EnglishFusion have a differential impact on the critical thinking skills of sub-groups of students?

The regression analysis has indicated that the only background variables that explain differences between groups are possession of classic literature, students' birth sex, ethnicity and parents' discussion of school performance. To explore the impact of the intervention on the CT skills of students with different characteristics, sub-group analyses were conducted according to demographic factors, home possessions, cultural capital, parental involvement, schools, prior academic and CT performance.

For brevity, this section presents only the results of the groups where there are interesting differences or where the difference is considerable. The detailed results are in Appendix H1. While the results of these sub-group analyses are interesting, they do not in any way explain the outcomes.

Comparison by demographic groups

Students were categorised into different groups based on age (younger and older, determined by the mean age), birth sex (boys and girls), and ethnicity (minority and majority). Students in the control group (mean age 13.99) were slightly younger than those in the experimental group (mean age 14.01). The age difference is small, but there is a perceptible difference in outcomes. Table 9.15 shows that the effect on younger students is bigger than for older students. This suggests that perhaps CT should be introduced at an earlier age, which is echoed by Kuhn’s (1999) proposition. Younger students exposed to EnglishFusion early could help them develop a more open and analytical mind and prevent biases and stereotypes. Older students, however, tend to be less adaptive to EnglishFusion. There is no differential effect on boys and girls. EnglishFusion benefitted both groups equally.

Table 9.15 Comparison of impact on gain in critical thinking by demographic characteristics*

Sub-groups	N	Effect size
Younger	986	+0.19
Older	1025	+0.08
Girls	1008	+0.16
Boys	1003	+0.12
Minority	39	+0.32
Majority/Han	1972	+0.13

* See Appendix H1 for more detailed results.

The effect on ethnic minority students’ CT skills appears to be considerably bigger than that for the Han majority. But this should be read with caution. The small number of minority students (n = 39) means that the results are very volatile to small changes.

Comparison by home possessions

Household possessions are used as an indicator of students' socioeconomic status (SES). In terms of household possessions, almost all students have a room of their own and a study desk (see 9.2 Characteristics of students). Comparing those with and without these possessions, there is no differential impact of EnglishFusion between groups. However, although the intervention has a positive effect on those with and without Wi-Fi at home, the intervention effect is bigger for those without.

Table 9.16 Comparison of impact on gain in critical thinking by household possessions

Sub-groups	N	Effect size
Without a room	119	+0.15
With a room	1892	+0.14
Without a study desk	207	+0.15
With a study desk	1804	+0.13
Without Wi-Fi	208	+0.28
With Wi-Fi	1803	+0.12
Without a computer	1531	+0.14
With a computer	480	+0.13

Comparison by children's cultural capital

Cultural capital refers to non-financial social assets that promote social mobility beyond economic means (Bourdieu, 1986). It is also considered as essential knowledge and skills needed to succeed in life (Office for Standards in Education, Children's Services and Skills [Ofsted], 2024). In the context of this study, it manifests through the possession of physical cultural objects, including classic literature, books of poetry, bookshelves, works of art, books on art or design and musical instruments.

The results showed that the intervention had a stronger effect on the CT skills of students who reported having classic literature at home than those who did not (see Table 9.17). Regression analysis shows that this variable explains a small amount of the variance in outcome.

Table 9.17 Comparison of impact on gain in critical thinking by cultural capital

Sub-groups	N	Effect size
Without classic literature	387	-0.08
With classic literature	1624	+0.18
Without books of poetry	872	+0.12
With books of poetry	1139	+0.15
Without a bookshelf	651	+0.14
With a bookshelf	1360	+0.13
Without works of art	1274	+0.14
With works of art	737	+0.12
Without books on art, music or design	1358	+0.16
With books on art, music or design	653	+0.10
Without musical instruments	1472	+0.18
With musical instruments	539	+0.03

However, the intervention has little differential effect for those who reported having books of poetry and those who did not. Having bookshelves and works of art at home does not accord any more advantage. Compared to the control group, experimental students who did not own books on art and design at home performed better than those who did. The effect is also stronger for those who did not have musical instruments compared to those who did. This may suggest that the intervention has an effect of overcoming disadvantages.

Comparison by levels of parental involvement in education

To estimate parental involvement in children's education, students were asked to indicate how often their parents did the five activities in the most recent academic year, assessing from 0 (never) to 10 (all the time). Based on the mean score of each activity, students were separated from being of a lower or higher degree of parental involvement.

It is hypothesised that children whose parents are more involved in their education may have an advantage, and perhaps have a higher level of CT. Regression analysis showed that of all the parental involvement variables, only the one about the frequency of parents discussing children's school performance is relevant. Students whose parents

are less likely to discuss their school performance benefit more from EnglishFusion than those who do not, suggesting that the intervention helps to overcome the disadvantage of parental involvement.

Table 9.18 Comparison of impact on gain in critical thinking by parental involvement

Sub-groups	N	Effect size
Low degree of discussing school performance	1095	+0.18
High degree of discussing school performance	916	+0.09
Low degree of helping with homework	1235	+0.13
High degree of helping with homework	776	+0.14
Low degree of discussing political or social issues	1072	+0.13
High degree of discussing political or social issues	939	+0.15
Low degree of going to a library or bookstore together	1149	+0.16
High degree of going to a library or bookstore together	862	+0.10
Low degree of discussing children's reading	1144	+0.14
High degree of discussing children's reading	867	+0.14

All the other factors do not contribute to explaining variation in the post-CT outcome. Children in both experimental and control groups with different levels of parental engagement performed similarly.

Comparison by schools

The intervention is effective in improving CT skills in all four schools (see Table 9.19). A marginally positive effect size is observed in schools B and D, and EnglishFusion is more effective for those from schools A and C. It is notable that schools C and D were small, each with only one experimental class. Any differences in students' CT skills could be attributed to teacher or class differences.

Table 9.19 Comparison of impact on gain in critical thinking by schools

Sub-groups	N	Effect size
School A	1132	+0.18
School B	618	+0.06
School C	153	+0.19

School D	108	+0.01
----------	-----	-------

Comparison by students' prior academic attainment

Students who scored lower than the mean score in the pre academic test were identified as lower academic achievers, while others were higher achievers. Table 9.20 shows that EnglishFusion had a stronger effect on higher achievers than lower achievers across all academic subjects. This is not surprising as high-achieving students typically possess strong cognitive abilities, allowing them to have a foundation for CT skills. Process evaluation (see Chapter 11) found that students with higher academic attainment were more engaged in EnglishFusion lessons while the less academically able students had greater difficulty understanding the CT instruction and were less interested and thus less engaged in class.

Table 9.20 Comparison of impact on gain in critical thinking by prior academic attainment

Sub-groups	N	Effect size
Lower achievers in Chinese subject	873	+0.02
Higher achievers in Chinese subject	1127	+0.25
Lower achievers in Maths subject	857	+0.13
Higher achievers in Maths subject	1141	+0.18
Lower achievers in English subject	946	+0.11
Higher achievers in English subject	1056	+0.22

Comparison by students' prior critical thinking skills

Of the 2,011 students, 907 pupils scoring below 7.70 (on a scale of 0 to 15) were categorised as lower critical thinkers, and 1104 pupils scoring above this threshold, were categorised as higher critical thinkers. Comparing the impact on lower and higher critical thinkers, the results showed that although EnglishFusion improved the CT skills of both groups, higher critical thinkers benefitted more than lower critical thinkers (see Table 9.21). In other words, those who were good at CT at the beginning made considerably more progress. This is not surprising as these students already have the foundation for CT, and the intervention capitalises on their existing abilities to apply their CT in the classroom.

Table 9.21 Comparison of impact on gain in critical thinking by prior critical thinking scores

Sub-groups	N	Effect size
Lower critical thinkers	907	+0.08
Higher critical thinkers	1104	+0.28

In summary, EnglishFusion is more effective for younger students, minority students, students from lower SES, higher academic achievers and higher critical thinkers.

9.6 Does training and teaching EnglishFusion alter teachers' critical awareness and attitudes towards teaching critical thinking?

Changes in teachers' critical awareness

To see if receiving training in CT infusion and having to deliver the lessons could also improve teachers' own CT, teachers were also asked to complete a simple five-question test, using a scale from 0 to 10. Before receiving the training, most teachers agreed that research conducted by well-known scientists, in reputable journals is trustworthy. They were also more likely to agree that studies that use standardised tests and large samples are trustworthy (see Table 9.22). The average ratings for these items are around 7.5 to 7.9 on an 11-point scale for the experimental group and 6.3 to 7.4 for the control. Teachers, however, are ambivalent about whether to trust research published recently (the average rating is 5.9 for experimental and 5.1 for control).

Table 9.22 Teachers' agreement on the trustworthiness of research findings (N = 21)

	Experimental teachers (n = 11)			Control teachers (n = 10)		
	Pre mean (SD)	Post	Gain	Pre	Post	Gain
1. Published recently	5.91 (3.05)	5.82 (2.52)	-0.09 (3.88)	5.10 (2.69)	6.50 (2.17)	1.40 (3.53)
2. Published in a reputable journal	7.45 (2.54)	7.36 (2.46)	-0.09 (3.18)	6.30 (3.65)	6.50 (2.12)	0.20 (2.39)
3. Conducted by a well-known scientist	7.55 (2.58)	6.18 (3.09)	-1.36 (4.41)	7.40 (2.99)	7.50 (1.72)	0.10 (1.73)

4. Supported by standardised test data	7.82 (2.32)	6.09 (2.12)	-1.73 (2.97)	6.90 (3.60)	6.80 (2.90)	-0.10 (2.23)
5. Conducted on a large sample size	7.91 (2.43)	7.45 (2.30)	-0.45 (3.80)	6.50 (2.27)	7.70 (2.87)	1.20 (2.39)

In comparison, the control teachers were more sceptical or critically aware than the experimental teachers at the beginning. Compared to experimental teachers, they were less likely to agree that recently published articles in reputable journals are trustworthy. Control teachers were also more sceptical about studies using the standardised tests and large samples.

However, after training and having delivered EnglishFusion lessons over three months, experimental teachers seemed to be more critically aware than the control teachers. They were less likely to agree on the trustworthiness of studies conducted by well-known scientists, or recent publications and those published in reputable journals.

Interestingly, the experimental group appears to be less trustworthy in research using standardised tests after the intervention than the control. The change is quite substantial. This could be related to the small sample size. The attitudinal changes of one teacher could considerably impact the overall results. For instance, one experimental teacher completely agreed with the use of standardised tests and large sample sizes at the outset but later somewhat disagreed. This is the vagary of a small sample size.

Changes in teachers' attitudes towards the teaching of critical thinking

To ensure the longevity of the intervention, it is important also to look at whether teachers' attitudes towards teaching CT has changed. For example, if teachers become more positive about the need to teach CT, they are more likely to continue to infuse explicit CT teaching in their lessons.

Teachers were asked if they thought CT should be taught in school, and whether it is relevant to the English curriculum because in China CT is less likely to be taught if it is not perceived as relevant. Both groups of teachers were very positive about the teaching of CT at the start of the study (see Table 9.23).

Table 9.23 Teachers' attitudes towards teaching critical thinking (N = 21)

	Experimental teachers (n = 11)			Control teachers (n = 10)		
	Pre mean (SD)	Post	Gain	Pre	Post	Gain
1. Should be taught in school	8.27 (1.42)	8.27 (1.90)	0.00 (2.32)	8.30 (2.83)	7.90 (2.18)	-0.40 (2.37)
2. Relevant to the English curriculum	9.36 (1.03)	9.18 (1.66)	-0.18 (1.60)	9.10 (1.60)	8.40 (2.32)	-0.70 (2.26)

Almost all agreed that it was relevant to the curriculum. At the end of the study (three months later), the experimental group was still very likely to agree that CT is relevant to the curriculum although there is a slight drop in rating from 9.36 to 9.18. However, the control teachers were even less likely to think that CT was relevant to the curriculum (a drop in rating from 9.10 to 8.40).

Overall, teachers in both groups were positive about the need to teach CT and its relevance in the curriculum. Although there was a slight drop in agreement at the end of the study, the decline was bigger among control teachers. For some reason, they were less likely to think that CT was necessary and relevant over time. Given the small number of teachers (n = 21), the results can be volatile. It only needs a couple of teachers to indicate differently to change the overall mean.

9.7 Chapter summary

The impact evaluation shows that it is feasible to infuse CT into the English curriculum in Chinese secondary schools and the results are promising. EnglishFusion has a small positive impact on students' CT skills, including making arguments, identifying assumptions, making deductions and inferences. However, the skills of interpretation do not appear to improve. This may be because EnglishFusion did not specifically address skills of interpretation. It could also be the case that this skill requires more time and exercises to be observable. EnglishFusion is found to be particularly beneficial for younger students, minority students, students from lower SES backgrounds, higher

academic achievers, and higher critical thinkers. Future studies could look at different groups of students to design tailored interventions for the improvement of CT.

The evaluation also suggests that EnglishFusion is not only helpful in improving students' CT skills but also effective in enhancing teachers' critical awareness. Teachers who have delivered the intervention hold more positive attitudes towards CT teaching at schools, and its relevance to the English curriculum in China. While this finding is tentative due to the small number of teachers ($n = 21$), it indicates the promising effect on intervention implementers who are often neglected in many trials.

Chapter 10 Impact of EnglishFusion on academic attainment

In addition to evaluating critical thinking (CT) skills, this study also examines the impact of the intervention on students' academic achievement. The previous chapter demonstrates that the infusion of CT into the English curriculum (referred to as EnglishFusion) appears to have a positive effect on students' CT skills. However, it is also essential to determine whether improvements in CT skills translate into enhanced academic performance, which is assessed through students' results in the three core subjects: Chinese, Maths, and English. This is particularly important, as academic outcomes are prioritised by schools in China. CT is unlikely to be emphasised in schools if its benefits for academic attainment are not evident.

This chapter addresses the following research questions:

1. Does EnglishFusion improve Chinese secondary students' academic performance?
2. Does EnglishFusion have a differential impact on the academic attainment of sub-groups of students (by age, birth sex, ethnicity, prior academic attainment, prior critical thinking skills, schools, parental involvement in children's education, and home background)?

10.1 Does EnglishFusion improve Chinese secondary students' academic performance?

Academic attainment was measured using the sum of the scores in the three core subjects: Chinese, Maths, and English language. These scores were obtained from students' final examinations that were administered across schools in the same district. This makes the test scores comparable across schools. Students' test papers were anonymously assessed by teachers to avoid bias. As previously noted, some cases had missing scores in the academic subjects (see Figure 9.1). Only students who provided academic scores were included in the analysis.

Students' prior academic attainment was compared to establish baseline equivalence. Because of the teacher-level randomisation across schools, there is some imbalance between the two groups (see Table 10.1).

Table 10.1 Comparison of pre-test academic scores between experimental and control groups

	Experimental group			Control group			Overall			ES
	N	Mean	SD	N	Mean	SD	N	Mean	SD	
Chinese	1018	78.35	18.33	1018	86.71	14.63	2036	82.53	17.10	-0.49
Maths	1016	63.95	31.86	1017	76.14	30.89	2033	70.05	31.96	-0.38
English language	1017	56.91	26.02	1020	71.26	26.11	2037	64.09	27.03	-0.53
Overall academic scores	1015	199.46	68.71	1014	234.43	64.81	2029	216.94	69.03	-0.51

At pre-test, experimental students lag behind their counterparts in overall academic attainment. Compared with the business-as-usual cohort, students involved in EnglishFusion achieved lower scores in English language (ES = -0.53), Chinese (ES = -0.49), and Maths (ES = -0.38). It is also clear that the English language was the weakest of all the three core subjects and Chinese being their strongest.

After the intervention, experimental students still appeared to have lower academic attainment in all three subjects, with the English language still the weakest (see Table 10.2).

Table 10.2 Comparison of post-test academic scores between experimental and control groups

	Experimental group			Control group			Overall			ES
	N	Mean	SD	N	Mean	SD	N	Mean	SD	
Chinese	1000	78.42	19.19	1006	86.95	14.74	2006	82.74	17.63	-0.48
Maths	999	60.07	29.36	1005	72.59	28.83	2004	66.33	29.79	-0.42
English language	1000	59.83	27.48	1007	74.76	28.04	2007	67.30	28.79	-0.52

Overall academic scores	992	198.43	67.74	1004	234.30	64.70	1996	216.37	68.62	-0.52
-------------------------	-----	--------	-------	------	--------	-------	------	--------	-------	-------

Given the imbalance in prior academic attainment, progress scores, rather than post-academic scores, were used to evaluate the impact of the intervention on academic achievement. The results suggest that the intervention had no beneficial effect on the growth of academic performance (see Table 10.3).

Table 10.3 Comparison of gains in academic scores between experimental and control groups

	Experimental group			Control group			Overall			ES
	N	Mean	SD	N	Mean	SD	N	Mean	SD	
Chinese	996	-0.31	10.99	1000	0.24	9.68	1996	-0.03	10.36	-0.05
Maths	994	-4.55	13.59	998	-3.76	13.75	1992	-4.15	13.67	-0.06
English language	996	2.43	10.57	1003	3.31	9.49	1999	2.87	10.05	-0.09
Overall academic scores	987	-2.36	21.96	994	-0.26	21.25	1981	-1.30	21.63	-0.10

The experimental group made slightly less progress than the control group across all three subjects (ES = -0.10), suggesting that there is no transfer of CT skills on academic performance in the short term. This is likely because the exams are very textbook-based, focusing on recall of factual information rather than critical analysis, evaluation or problem-solving. In other words, they test different skills. The intervention itself does not address the initial low academic performance.

Sensitivity analysis

As there were missing cases from both experimental and control groups (see Figure 9.1), results may be biased. For example, if those missing the post-test from the experimental group were high performers, excluding them from the final analysis may dampen the overall effects and vice versa if most of the control students missing the post-test were high performers. To test how sensitive the overall effects are as a result

of the missing cases, Number Needed to Disturb (NNTD) analysis was calculated (see Table 10.4). The larger the NNTD, the more secure or stable is the result because this means that you will need that number of cases to have opposite effects to alter the results.

Table 10.4 Results of sensitivity analysis of academic attainment

	Smaller cell	Missing cases	Effect size	NNTD
Chinese	996	50	-0.05	50
Maths	994	54	-0.06	60
English language	996	47	-0.09	90
Overall academic scores	987	65	-0.10	99

For the Chinese subject, there were 50 missing cases in total (26 from the experimental group and 24 from the control group). NNTD is calculated as the effect size multiplied by the size of the smaller group in the comparison, which is $0.05 * 996 = 49.8$ (roughly 50). This means that all 50 missing cases would need to have opposite effects to alter the overall results. As there are exactly 50 missing cases, there is a potential for the results to change if these cases had reverse effects.

The result for Maths is slightly more stable, with an NNTD of 60 compared to 54 missing cases. This means that even if all 54 missing cases had opposite effects, it would not be sufficient to change the overall effects.

For the English language subject, the result is also more robust due to the smaller number of missing cases compared to the NNTD. The NNTD for English language is almost double the number of students who did not take the final examination. This means that the result is less likely to be affected by the English missing data.

The overall results can be considered secure as NNTD is bigger than the number of missing cases. In summary, attrition did not unduly influence the substantive results in English and overall academic scores, but the results of the intervention's impact on gain scores in Chinese and Maths should be treated with caution.

A comparison of the pre-test academic scores between students who missed the post-test in the experimental and control groups can offer further insights into the intervention’s impact on academic attainment. Table 10.5 shows that, among students without post-test academic scores, those in the experimental group had considerably lower pre-test academic achievement compared to those in the control group. This suggests that the experimental students who dropped out were generally lower academic achievers. If these students had been included in the post-test, the overall impact of the intervention may have been more negatively affected.

Table 10.5 Comparison of pre-test academic scores of students missing the post-tests

	Experimental group			Control group			Overall			ES
	N	Mean	SD	N	Mean	SD	N	Mean	SD	
Chinese	22	58.91	26.40	18	76.17	17.46	40	66.68	24.16	-0.71
Maths	23	36.64	26.33	19	47.26	28.75	42	41.56	27.65	-0.38
English language	22	37.38	22.57	17	49.59	18.36	39	42.84	21.42	-0.57

Additional analysis

In addition, a multivariate regression model was developed to determine the extent to which the academic outcomes could be attributed to the intervention. The post-test total academic score is the dependent variable. The potential influencing variables were entered in blocks, with background factors (i.e. age, birth sex, ethnicity, SES, and parental involvement) entered first. Students’ prior academic scores were entered next, followed by pre-test CT scores. Characteristics of schools and teachers were considered in the next block. Membership of the treatment group was added in the final block to see how much the intervention explained the results after controlling for all the other factors.

Using the forward method, variables with minimal predictive power were excluded to construct a model that maintains the R-squared while reducing the number of factors. Of all the household items used as a proxy of students’ home background, possession of classical literature and books of poetry are relevant (see Table 10.6). Similarly, of all parental involvement activities, only discussing school performance and going to the

library or a bookstore are relevant. Overall, excluded variables include age, ethnicity, other household items such as a room and a computer, and other parental involvement activities in children’s education.

Table 10.6 Regression results predicting students’ post-test academic scores

	R	R Square	R square change
Classic literature, discussing school performance, books of poetry, birth sex, going to the library	0.27	0.07	0.07
Pre-test academic scores	0.95	0.90	0.83
Pre-test CT scores	0.95	0.90	0.00
Teacher characteristics	0.95	0.90	0.00
Intervention	0.95	0.90	0.00

Students’ background factors explain only 7% of the variance in post academic outcomes (see Table 10.6). Adding their prior academic score to the regression alone explains an additional 83%, making it the strongest predictor. Since the post academic test was conducted only three months later, students’ post-test scores are likely to be closely related to their pre-test performance. Teacher characteristics do not add to explaining students’ post-academic outcomes. The intervention makes no difference. Future studies should look also at changes in academic outcomes over a longer period rather than in the same school term. The overall model correlates strongly with overall academic outcomes with an R of 0.95 and R^2 of 0.90 accounting for over 90% of the variation in post academic outcomes. This is consistent with the impact evaluation, which shows that the intervention did not improve students’ academic outcomes. The results of the regression show that the strongest predictor of students’ academic outcomes is their prior performance.

The coefficients for each variable reveal the strength and direction of the correlation to post academic attainment. As the predictors were measured using different units and scales, the standardised beta coefficients were used. The standardised coefficient of students’ prior academic attainment (+0.94) suggests that it is the best predictor of the post-test score in academic subjects (see Table 10.7). Other variables, such as the

teacher characteristics and whether students receive the intervention or not are not important in explaining the academic outcomes. This means that once students' prior academic scores are accounted for, the teacher and intervention make no difference to students' post academic attainment.

Table 10.7 Coefficients for the model predicting post-test academic scores

	Unstandardised <i>B</i> coefficient	Standardised beta coefficient
Constant	9.02	-
Classic literature	3.38	0.02
Discussing school performance	0.01	0.00
Books of poetry	-0.72	-0.01
Birth sex	-1.76	-0.01
Going to the library	-0.03	0.00
Pre-test academic scores	0.94	0.94
Pre-test CT scores	0.19	0.01
Teacher characteristics	0.18	0.02
Intervention	-3.51	-0.03

10.2 Does EnglishFusion have a differential impact on the academic attainment of sub-groups of students?

Similar to the sub-group analyses for the primary outcome (CT skills), analyses were conducted to compare the impact of the intervention on the academic outcomes of various sub-groups: age (younger or older), birth sex (girls or boys), ethnicity (minority or majority), socioeconomic status (SES, with or without a series of household items), parental involvement in education (higher or lower engagement), schools, prior academic scores and CT skills. Only students who provided complete gain scores for academic achievement were included in the analyses. For brevity, this section presents only the results for the groups where the differences are substantial. Detailed results for all sub-groups are in Appendix H2.

Comparison by demographic groups

Students aged under 14 years old were categorised as younger students, while those aged 14 and older were classified as older students. Neither group appeared to benefit from the intervention in terms of improving their academic performance (see Table 10.8). However, EnglishFusion appears to have a stronger negative effect on older students' academic outcomes. It also has a stronger negative impact on girls' academic outcomes than for boys.

Table 10.8 Comparison of impact on gain in academic scores by demographic characteristics*

Sub-groups	N	Effect size
Younger	974	-0.06
Older	1007	-0.13
Girls	996	-0.15
Boys	985	-0.05
Minority	39	-0.08
Majority/Han	1942	-0.10

* See Appendix H2 for more detailed results.

Comparison by home possessions

The results are interesting. It appears that EnglishFusion had a particularly beneficial impact on students with fewer household items. Compared to control students who had their own room, study desk and Wi-Fi, experimental students in this group made less progress in academic outcomes than those in the experimental group who lacked these objects (see Table 10.9). This perhaps suggests that the intervention had some effect on addressing family disadvantage.

Table 10.9 Comparison of impact on gain in academic scores by home possessions

Sub-groups	N	Effect size
Without a room	119	+0.06
With a room	1862	-0.11
Without a study desk	199	+0.02
With a study desk	1782	-0.11

Without Wi-Fi	202	+0.04
With Wi-Fi	1779	-0.11

Comparison by children’s cultural capital

Comparing the impact of the intervention on academic attainment of students with different levels of cultural assets (i.e. possessions of classic literature, books of poetry, bookshelves, works of art, books on art or design and musical instruments at home), the results showed that compared to control students, experimental students who reported not having classic literature, books of poetry and works of art at home, made considerably less progress than those who did. This indicates that the intervention did not help to address the disadvantage of lack of cultural capital at home.

Table 10.10 Comparison of impact on gain in academic scores by cultural capital

Sub-groups	N	Effect size
Without classic literature	379	-0.12
With classic literature	1602	-0.09
Without works of art	1254	-0.12
With works of art	727	-0.07
Without books of poetry	855	-0.15
With books of poetry	1126	-0.06

Comparison by parental involvement

The regression analysis result shows that of all parental involvement activities, only discussing school performance and going to the library or a bookstore are relevant to academic outcomes. Students whose parents less frequently discussed school performance with them showed greater negative changes in academic scores (see Table 10.11). These students have a lower academic background (pre mean = 210.62) than those with higher parental involvement (pre mean = 228.17) in this respect. It may take them a longer time to enhance academic outcomes. There is no differential impact for groups with different degrees of going to a library or bookstore.

Table 10.11 Comparison of impact on gain in academic scores by parental involvement

Sub-groups	N	Effect size
Low degree of discussing school performance	1073	-0.15
High degree of discussing school performance	908	-0.03
Low degree of going to a library or bookstore together	1132	-0.09
High degree of going to a library or bookstore together	849	-0.11
Low degree of discussing political or social issues	1057	-0.02
High degree of discussing political or social issues	924	-0.18
Low degree of discussing children's reading	1122	-0.07
High degree of discussing children's reading	859	-0.14

Interestingly, the intervention appears to have a stronger negative effect on those with a higher degree of parental involvement in discussing political or social issues and children's reading. While these students may have the foundation for CT as engaging in political discussions encourages students to be critical, evaluate evidence and construct reasoned arguments, this is likely because the academic exams focus on recall of factual information taught by the textbook.

Comparison by schools

EnglishFusion appears to be effective in improving students' academic scores in all schools except for School A (see Table 10.12).

Table 10.12 Comparison of impact on gain in academic scores by schools

Sub-groups	N	Effect size
School A	1113	-0.24
School B	615	+0.14
School C	147	+0.13
School D	106	+0.22

Around half of the participants were from School A, meaning the overall effect size may be pulled down by the circumstances at School A. School A had the largest difference in prior academic scores between experimental and control groups, with the experimental group having much lower scores than their counterparts. This imbalance

suggests that students from School A were not academically balanced after randomisation, likely due to class segregation practices in this school. Students with higher academic scores were often assigned to the same class, while other classes comprised mediocre students. Although this study involved the randomisation process, it was at the cluster/teacher level. It could not avoid the pre-existing academic score imbalances between classes.

The intervention had no differential impact on other sub-groups (see Table 10.13).

Table 10.13 Comparison of impact on gain in academic scores by other sub-groups

Sub-groups	N	Effect size
Lower academic achievers	877	-0.12
Higher academic achievers	1104	-0.19
Lower critical thinkers	887	-0.08
Higher critical thinkers	1094	-0.11

In summary, the findings suggest that the intervention has a smaller negative effect on the academic attainment of younger students, boys and students with a lower parental involvement. EnglishFusion is more effective for students from lower SES backgrounds.

10.3 Chapter summary

While EnglishFusion is effective in enhancing students' CT skills, it has no beneficial effect on their academic attainment. This means that there is no transfer of CT skills to academic performance. It is still students' prior academic performance that largely explains their post academic outcomes. Despite the overall negative effect size (-0.10), it is interesting that students from Schools B, C and D benefited to improve their academic scores.

In conclusion, there is no evidence that EnglishFusion supports the improvement of students' academic achievement. The dropout cases and initial imbalances in academic performance between experimental and control groups may influence the evaluation. The potential bias of class segregation, where students with similar academic performance are grouped together, could also have affected the outcomes, given that randomisation occurred at the teacher level. Future studies should strive to ensure a

balance between experimental and control groups to better assess the intervention's effectiveness.

Chapter 11 Results of process evaluation

The preceding chapters present the results of the impact evaluations. The results suggest that EnglishFusion has the potential to enhance students' critical thinking (CT) skills, but it does not have a similar effect on academic attainment. A process evaluation was conducted to help explain the results and the mechanisms of change. It helps to assess fidelity of implementation, that is, whether the teachers delivered EnglishFusion as they have been trained and as per the lesson plans developed for delivery. Data was collected from interviews with both experimental students and teachers and lesson observations of both experimental and control groups. Conditions and challenges for the successful implementation of EnglishFusion are also summarised.

11.1 Fidelity to implementation

On the whole the intervention was implemented as intended. Experimental teachers successfully maintained the confidentiality of materials from control teachers during the intervention period to reduce the risk of diffusion. Control teachers primarily focused on the formal English textbook and tests, employing a teacher-centred approach that concentrated on delivering English language knowledge.

In addition to teacher instruction and explanation, students were presented with slides for self-reading and comprehension. All teachers demonstrated use of these slides in their classes except for one teacher during one lesson due to a power outage. The teacher compensated by writing key notes on the blackboard before class and providing handouts for students to read and write during the lesson, thus minimally impacting the delivery of the intervention.

It is worth noting that lesson plans did not mandate strict adherence but provided general guidance on CT instruction. Teachers were granted autonomy in delivery methods, allowing flexibility to accommodate student needs. For instance, activities could involve discussion, individual idea generation, or a combination thereof, tailored to student characteristics. Teachers could also revise slides with additional translations as needed, enhancing comprehension. These implementation variabilities aimed to optimise intervention appropriateness for the student group and foster CT development.

11.2 Teacher training and preparation

Prior to the implementation of EnglishFusion, experimental teachers attended a formal training session. This was conducted on 2nd March 2023. All 11 experimental teachers attended the training, which focused on the practical application of the infusion CT teaching method. The teachers were engaged and demonstrated great interest in the training. Initially, I conducted a demonstration of the second EnglishFusion lesson, which introduced the concepts of relevance and the straw man fallacy, allowing the teachers to experience the lesson from a student's perspective. Additionally, teachers were organised into groups to practise teaching the first and third EnglishFusion lessons, with the opportunity to suggest more answers than those in the lesson plan. During this practice, concerns about language and class discipline were raised. For instance, in exercises requiring students to interpret multiple meanings of a sentence, teachers felt that direct translation from English to Chinese might obscure some meanings. They were also concerned about maintaining class discipline if the class became too noisy. The importance of providing feedback to students' answers was also highlighted. After asking open-ended questions, teachers needed to evaluate the appropriateness of student responses. These concerns demonstrate that teachers had thoroughly considered EnglishFusion content and teaching methods.

Informal weekly follow-up training sessions were also conducted. As teachers gained experience with EnglishFusion lessons, they became more comfortable with the infusion method and were more confident in suggesting improvements. For example, they suggested removing the topic on school trips from the third lesson, as their students lacked experience in this area. They also recommended rewarding debate competition winners, such as reduced homework or bonus points. Additionally, some teachers scheduled to teach EnglishFusion later in the week benefited from knowing potential issues in advance. For example, if students were not engaged in group activities, the experimental teacher and I would discuss potential reasons and provide practical solutions for other teachers. Based on class observations, I also provided reminders about possible challenges in the delivery of the intervention, such as insufficient time to complete all planned activities, the teachers' ability to ask prompt questions, and their tendency to over-explain for students.

Most teachers were conscientious in preparing for EnglishFusion lessons. Teachers from the same school often formed groups to collaboratively prepare, ensuring they understood the content and were familiar with the teaching process. They regularly contacted me for further clarification if they had any questions or disagreements.

However, transitioning from a teacher-centred method to the CT teaching approach takes time. Some teachers were inclined to provide answers directly and over-explain. In some cases, students were given limited time for thinking and discussion. Classroom observations indicated that one or two teachers were not fully familiar with the lesson content and delivery procedures, especially towards the end of the first stage. These teachers relied heavily on printed lesson plans, leading to repetitive instructions and unresponsive student reactions. Students felt bored and confused when teachers were overly dependent on the lesson plans, and some complained about the lack of clarity in the teaching.

Overall, the teacher training and preparation were successful. The experimental teachers involved in the primary research had no prior experience of CT teaching. At first, they thought EnglishFusion lessons would be abstract and complicated. Some of them had only heard of the term “critical thinking” but had no clear understanding of it. As one teacher expressed:

I was confused (about the term critical thinking). What is this? How am I supposed to teach it if I don't understand? It was really confusing at first.

However, after the teacher training and the progression of EnglishFusion lessons, they found it much easier to teach and became more familiar with the infusion method.

11.3 Students' opinions on EnglishFusion

Lesson content

Most students were new to CT lessons. Initially, they thought EnglishFusion was a lesson on solving abstract problems. For example, one student said:

It was like giving you puzzles and then asking you to complete it, and teachers would teach complicated and abstract ideas that students do not understand.

However, after experiencing EnglishFusion lessons and tasks, they realised it was different from what they had imagined.

It turns out that this is different. The teacher is leading us and using multimedia, like those videos, and then giving us a message of what is going on, so that we can judge the information ourselves. It is interesting to discuss it with classmates.

It turns out to be not as difficult as I thought it would be after learning (EnglishFusion lessons). I was able to participate very well. In fact, sometimes I was able to answer some of the questions raised by the teacher.

I was surprised by this EnglishFusion course. I thought it might just be about asking us to practice more test questions, but this was not true. Thinking element was infused, and we were asked to learn by ourselves, and embraced a new way of learning.

Students found the intervention interesting because the examples and exercises were closely related to their daily lives, allowing them to easily connect EnglishFusion content to their prior experiences and knowledge. As one student commented:

EnglishFusion is interesting. We need to use our own life experience to think about the examples. Otherwise, we cannot get the answers. (EnglishFusion) is attractive, and it allow us to think deeply, and provide interesting answers.

However, some students preferred to be presented with unfamiliar materials. One student noted that discussing novel issues in class was more engaging than talking about

familiar ones. Another student mentioned their interest in the UFO material because it was new to them: “I felt fortunate to (get to know that material) and was concentrated in that lesson.”

Students were generally capable of understanding most of EnglishFusion content. They could identify straw man arguments in communication and knew how to argue against them. Some found it easy to apply in practice, improving their communication skills with others. They also became more discerning about the information they received. As one student said:

We often read some messages on the Internet. We can use what we have learned from EnglishFusion lessons to evaluate the trustworthiness. We do not believe some rumours. We can judge the information.

Nonetheless, not all lesson content was considered easy. Some students found it challenging to learn about correlation and causation and making inferences. They found these lessons difficult because the examples were more abstract. Additionally, in these lessons, students were asked to answer close-ended questions with only one correct answer. This allowed them to assess whether they had truly understood EnglishFusion content, but some were worried about being laughed at when they selected a wrong answer. Overall, most students indicated that the first few EnglishFusion lessons were easy, while the latter part became slightly more difficult.

Lesson delivery

The format of delivering EnglishFusion also appealed to students. They enjoyed the use of pictures, videos, and slides, which contrasted with the regular curriculum that relied heavily on textbooks. Some students found the textbook-based learning boring. More pictures and videos were suggested to enhance engagement. EnglishFusion lessons slides were perceived to be more vivid than those used in usual classes. As one student explained:

I think EnglishFusion is interesting because the teacher shows us some pictures, videos, and stories, where we can find out clues to complete EnglishFusion activities.

Students also appreciated the group discussions. They had many opportunities to express their ideas and listen to others' opinions. Group competitions were also set up to encourage greater engagement. They particularly enjoyed the third lesson, which involved discussing the advantages and disadvantages of shorter school days. Here are some student perspectives on the value of peer discussions:

I think it is quite interesting. Sometimes when I discussed with my classmates, we had different views. I felt a great sense of achievement when I successfully convinced them.

When other students' views were different from mine, we might come up with some new ideas by debating. It was very interesting, and it made us feel that our thinking skills improved.

Improvement in student-teacher relationship

Students found it fascinating that teachers were not as serious as usual. The teaching style shifted to being student-centred, with teachers becoming more open and active. Some students noticed the difference in pedagogy between the regular class and EnglishFusion lessons:

EnglishFusion course is somewhat different from the traditional way of teaching. Traditional education is teacher-centred, and it is one-sided input for students. However, EnglishFusion teaching is a two-way communication. It is a mutual cooperation that searches for the truth.

Students felt they had a closer relationship with teachers after EnglishFusion lessons. When they disagreed with the teacher, they were more willing to point it out and explain

their stance, fostering a comfortable classroom atmosphere where students could express, justify, and share their ideas. One student recounted:

Student: I did not know (how to think from different perspectives) before, but when the teachers said something in class, I was able to point out the mistake in their arguments.

I: What is the teachers' reaction after you pointed it out?

Student: They support (my point).

Improvement in student confidence

EnglishFusion lessons were perceived to be not only interesting but also relaxing. There was less work assigned to students, with no need to memorise new vocabulary or grammar, and no homework. Most questions were open-ended, allowing students to provide their own answers without worrying about being criticised for wrong answers. For example, some said:

(EnglishFusion lessons) are very relaxing and enjoyable, allowing us to say what we want to say without worrying about getting it wrong.

EnglishFusion course is friendly to students like me whose English was not very good. We can infer the meaning of new words based on what we have learned from the CT course. If we meet new words, there are translations. I would always take notes and sometimes review them to expand my knowledge scope. I felt a little proud of myself after all these things were done. It levels up my confidence.

Overall, the intervention was well-received and considered both interesting and relaxing. As one student expressed:

To be honest, I look forward to attending EnglishFusion lesson every week. The lesson content is interesting, and it is like doing extracurricular reading exercises. It is a pity that the seven lessons

pass so quickly. I hope there will be more (EnglishFusion lessons) in the future. I really like these lessons that could open my mind.

11.4 Teachers' views on EnglishFusion

Lesson content

EnglishFusion content was believed to become gradually more difficult. The first few lessons were considered easy and appropriate for their students, but the lesson on making logical inferences by drawing diagrams was more challenging. Despite this, most students could understand the content with teachers' guidance.

Teacher: The difficulty level of these examples and exercises is basically in line with our eighth-grade students' cognitive ability. They are close to life. I think you chose them carefully. Otherwise, if they were too abstract, student engagement would definitely not be that high.

I: Does this mean that most of the students in the class can understand the examples and exercises?

Teacher: Yes, most of them. I suppose (the participation rate) could reach two-thirds.

Improvement in student engagement

All teachers liked EnglishFusion content and believed it was interesting and interactive. They noticed that EnglishFusion was different from the usual English course. Some teachers commented:

The usual English class was more serious. Sometimes students felt (the English course content) was too abstract to understand, but we teachers had no choice. We had to finish our teaching task, so I had to cram knowledge into them regardless of whether they understood it. This is spoon-feeding education.

EnglishFusion was different. Our usual English course is just about memorising vocabulary and sentences, including listening, speaking,

reading, and writing. Generally, this is how we English teachers at secondary schools teach.

Compared to the usual English course, teachers found that students were more active and engaged in EnglishFusion lessons. One teacher anticipated this might be because the class atmosphere was more relaxed, and students could use Chinese rather than English to express their ideas.

Students were required to answer questions in English in the usual English course. They were more active and courageous in sharing ideas when they were allowed to use Chinese in EnglishFusion course. They looked forward to having EnglishFusion lesson every week.

Another teacher attributed the increased engagement to the opportunity for group discussions and the attractiveness of the course content and slides.

I really like this (EnglishFusion) course. It allows students to express ideas freely. It is different from our usual class where the teacher leads students to learn. (EnglishFusion) allows students to think more autonomously. Group activities are mainly used to stimulate students' enthusiasm. The students' ability to communicate with each other has also increased greatly. In addition, the course slides and examples are very appropriate, and I think students are also very interested.

11.5 Perceptions of the impact of EnglishFusion on critical thinking skills

Most students were convinced that their CT skills had improved through EnglishFusion. Students noted that they could now better assess the relevance of conversations, evaluate information, reflect on common assumptions, and distinguish between correlation and causation. They provided specific examples to illustrate how EnglishFusion had positively impacted their thinking.

After learning EnglishFusion course, I think my thinking skills have been developed. Take the straw man fallacy as an example.

Sometimes when a classmate was talking, I always had no idea what he/she was talking about. Now I can understand why I did not know that.

The one I have used widely is judging information. I read some news saying that China was providing (weapons) to Russia and, at the same time, providing (weapons) to the Ukrainian army. I realised that it was a rumour after checking for other information. I argued against the author (of the news), and that person directly deleted my comments.

I used to accept all arguments without thinking about them carefully. It did not matter for me to think whether these arguments were correct. People said that girls were born to be good at liberal arts, while boys were bad at them, and good at science. When you play the phone, you would have poor eyesight, and then your academic grades must decrease. It seems that these results are of 100% certainty. Previously, I did not care about them, and it did not look like it affected me. But this time, I think it is different, and I think they are wrong. These factors and results are just correlated, why do people assume there is a causal relationship?

All teachers also agreed that EnglishFusion had improved their students' CT skills. One teacher provided an example of students evaluating the trustworthiness of a video in their regular English class:

Teacher: I think (EnglishFusion course) had an influence (on students' thinking). A good example was when I played a video for them yesterday. My intention was for them to watch the video and learn about its content. But all of them said it was fake. They were judging the credibility of the video. Even when I told them it was true, they still said it was fake. I think it really affected their thinking. Now, when I hand out course materials, they judge whether the material is accurate rather than (directly) analysing it.

I: If they directly gave the conclusion, have you asked them how they arrived at it?

Teacher: I did not ask them. They consciously gave reasons why (they thought) it was fake. They said, “Look at the video. The person looks fake, and the building suddenly disappears. It is made of three-dimensional animation.”

Other teachers assessed the influence on students’ thinking skills based on their preparation for English tests. Some students could apply what they had learned from EnglishFusion lessons to English practices, such as writing compositions and reading articles. They could think from different perspectives. For instance, before learning EnglishFusion lessons, students could only write a few sentences based on the question prompts. However, afterward, they were able to write more through divergent thinking. In reading exercises, they could discern differences between similar options. Additionally, after exposure to the intervention, students were reported to provide explanations or examples to validate their arguments, which differed from their previous responses of just “yes” or “no”.

It should be noted that almost all teachers believed students with average academic scores and above were more receptive to the CT courses. These students attended EnglishFusion lessons more seriously and were more engaged. As one teacher commented:

(EnglishFusion) has a positive effect on students with good grades. It is also beneficial for students with average academic grades. However, I think that students who are academically behind might not understand at all. I could tell this from their facial expressions.

This observation aligns with the sub-group analysis of the impact evaluation, which found that higher academic achievers made better progress in their CT skills.

11.6 Perceptions of the impact of EnglishFusion on academic attainment

Students' perceptions corroborated the impact evaluation results, which found no impact on academic attainment. Students had mixed opinions on the influence of EnglishFusion on their academic learning, particularly regarding English language acquisition. Several students believed that EnglishFusion could enhance their English learning. They felt it encouraged them to seek alternative answers, improve writing skills, analyse textbook articles, expand vocabulary, and increase their enthusiasm for learning English. Their judgements were often based on performance in English tests, such as reading exercises and writing practices. Some students also mentioned that EnglishFusion could help them achieve higher scores in examinations for Chinese and Maths.

However, a larger number of students believed their academic learning, particularly their English learning, was not influenced by the intervention. They said:

I do not think it had much of an impact. Current English learning requires comprehension, memorisation, and recitation. It had little to do with thinking.

EnglishFusion lessons had little impact on my English. I can make fluent sentences and write new words in my notebook. But I did not memorise them, and it felt upsetting. I did not get a bonus score if I included them in the writing composition.

Some students acknowledged that they did not know how to apply EnglishFusion content to their regular studies. This may explain why the positive impact on CT skills does not extend to academic attainment. Another reason for the ineffective impact was the perception that EnglishFusion was not relevant to their academic studies, leading to lower engagement levels in class. As the following conversation shows,

Student: I found that two-thirds of my classmates were very engaged in (EnglishFusion) class. The rest were not very active.

I: What might be the reason?

Student: One is that some students are not very interested. For example, my classmate was sleeping (in EnglishFusion lessons). Alternatively, some students were not good at English, and they might be less interested. They did not understand the content.

Another student: I agreed with this. I found some students were sleeping in (EnglishFusion) class. Besides, even in my group, some of them discussed irrelevant issues. It had nothing to do with their abilities. The reason is that they did not think EnglishFusion content was relevant to them. There were many new words they did not use. Even if the teacher had explained, it cannot be memorised completely.

Teachers were also hesitant to assert that there was a positive impact on students' English learning. While some teachers mentioned that EnglishFusion could help students understand articles better and write better compositions, most thought the influence, if any, was minimal. One major reason was that EnglishFusion was not designed for English tests. According to one teacher,

A large part of our (English) test involves (close-ended) reading questions. If students use what they have learned from EnglishFusion lessons, it is difficult for them to choose the correct answers.

This echoes students' views that EnglishFusion content was not tested in their academic tests. They acknowledged that while EnglishFusion might not directly improve their test scores, it would be useful for those planning to study overseas. Examples from the student interviews:

If I am going to study abroad, I think EnglishFusion is very useful. However, I think it is not very useful in the regular examinations. This is because English learning is just limited to words, phrases, reading, and comprehension and so on in China. For me, it may not be useful (in the English tests), but it enables us to open our minds and understand what others talk about when we are abroad.

EnglishFusion course did not improve our academic scores. It did not help a lot. If we are going to study abroad, it is more helpful in communication and writing English essays. But it is not useful for the English course like we have at school. We think EnglishFusion is fun, and we are eager to know about it. It is like playing games.

Additionally, some teachers thought it would take a longer time for the influence on students' English learning to become observable. This suggests that the intervention might need to be implemented for a longer duration. Moreover, since Chinese language was primarily spoken in EnglishFusion lessons, some teachers believed there was not much influence on their students' English proficiency.

11.7 Challenges to successful implementation

Limited teaching time

One major problem that English teachers at secondary schools in China face is limited teaching time. According to the curriculum plan for compulsory education in China (MoE, 2022a), the proportion of English teaching hours within the total teaching hours (6%-8%) is much smaller than that for Chinese (20%-22%) and Mathematics (13%-15%). Some teachers argued that English teaching hours are even fewer than those allocated for art subjects. Consequently, teachers have to complete the regular English curriculum within this restricted timeframe. Although EnglishFusion is more time-efficient than setting up a new curriculum, it still requires adequate time for classroom instruction (Solon, 2007). For example, some teachers reported that they could not complete the planned EnglishFusion content because they allocated too much time for discussions and lost track of time while listening to many groups' ideas.

Heavy workload of teachers

Another challenge is the heavy workload faced by English teachers. They are required to cover a textbook that usually contains ten units. Additionally, many teachers are carrying out various irrelevant teaching issues such as administrative activities (Liu & Onwuegbuzie, 2012; Yan, 2015). Infusing CT into their classes would increase their preparation workload, adding to their already substantial responsibilities.

Teaching to the test

The emphasis on teaching to the test also hinders the successful implementation of the CT course. Most teachers and students are primarily concerned about English tests. Teachers must ensure that the content to be tested at the end of each term is thoroughly taught. Otherwise, their teaching is considered incomplete. Despite the limited teaching hours allocated to English (MoE, 2022a), the English test score remains a critical component of the senior high school entrance examination. This creates a major challenge for English teachers, as they must prepare students for tests within a short period. To efficiently cover the required content, teachers often resort to a teacher-centred approach, directly providing answers and explanations to students. This method allows teachers to manage the class effectively and cover as much knowledge as possible.

Meanwhile, students tend to neglect knowledge or skills that are not tested. For example, when asked about methods for collecting information on the number of eggs people consume per day, only one or two students suggested using random sampling. Some students said that although they had learned about sampling in their Maths class, they did not pay attention to it because it was not tested. This indicates that if a topic is not tested, it is unlikely to be taught in class.

Students' poor English skill

Another obstacle is the lower proficiency of students in English. Teachers have indicated that most students did not begin learning English systematically until secondary school. Limited vocabulary and grammar knowledge hinder their understanding of CT content. Without translations, students often have no idea what to discuss. Additionally, if the course content is presented in long English texts, students are reluctant to read it.

A lower level of English proficiency could prevent some students from understanding the lesson content. As one student said in the interview:

I: You did not understand that (EnglishFusion) lesson either?

Student: I did not understand it at all because I am very bad at English.

I: Is it the difficulty of the English language that makes you less inclined to listen to the lesson?

Student: It is not that I do not want to listen to it. I just do not understand it. I want to follow EnglishFusion lessons carefully, but when I look at the new English words, I do not know the meaning. I become really nervous, particularly when the teacher asked me (to answer questions).

Large class size

Large class sizes (50-60 students per class) pose another challenge. When one or two students provide the correct answer, teachers tend to move on to the next exercise. Students who can answer correctly are usually those who understand the lesson content, while those who are confused are less likely to speak up. Teachers cannot attend to all students in such large classes, making it difficult to provide timely support. A smaller class size would allow teachers to better oversee all students and offer assistance to those struggling.

Additionally, some students are uncomfortable speaking in front of large groups. They may fear giving wrong answers and being laughed at. For example, some students chose option B for a question but did not respond when asked by the teacher. However, when they saw that most students raised their hands for option C, they followed their peers. This behaviour highlights the reluctance of some students to participate actively in large classes.

11.8 Conditions for successful implementation

Comfortable class atmosphere

Teachers are responsible for creating a comfortable classroom atmosphere to foster CT. When students are asked to answer questions, they should be given sufficient time and multiple opportunities to discuss with peers. The course needs to be delivered in a relaxed manner. Specifically, students should not be burdened with excessive workloads such as memorising and taking notes. If they miss part of the course, they should not worry about falling behind classmates or receiving lower test scores. This approach will boost their confidence in expressing their ideas, as they will not be

criticised for providing different answers. Additionally, a comfortable class atmosphere requires teachers to provide feedback on students' responses. This interactive process helps students receive timely responses and better understand the teacher's views. When competitions are held, offering attractive rewards for winners is advisable. This strategy can help maintain students' focus on the CT course and provide them with a sense of achievement upon receiving rewards.

Teachers being supported

Another factor that facilitates the integration of CT teaching is the support provided to teachers. According to teacher interviews, they suggested that they should receive training to teach CT effectively. If CT is to be infused into English teaching, teachers need to be well-prepared. They require more time and resources to infuse CT effectively into the English classroom. Practical training is essential to help them become familiar with CT development, improve their skills in asking thought-provoking questions, and enhance their ability to translate and explain concepts clearly.

Closely link to textbooks and tests

If CT teaching is closely linked to the textbook and tests, teachers will be more inclined to teach the course, and students will place greater importance on it. The combination of CT, textbook knowledge, and tests is highly recommended by both teachers and students. Teachers explained that they were under a heavy workload to complete teaching the textbook content, but there was limited time. It would be a waste of time if the English textbook was not used.

Students with better English skill and more background knowledge

Surprisingly, all teachers anticipated that EnglishFusion might have a better influence if it is taught in more advanced areas. They maintained that their students were not proficient in English and students from urban areas were perceived to have more access to educational resources and thus have more background knowledge. As one teacher explained,

First of all, children in big cities have more knowledge on a wider range of topics, so they can better understand what you are talking

about. Second, they are more adaptive to the English language of the CT course. Besides, they can generate more ideas. So I think the teaching results will be better if the course is taught in a better place.

This aligns with the sub-group analysis result that students with higher English skills are more receptive to EnglishFusion. These students tend to be more engaged and understand the lesson content better. Some of them are confident in expressing their ideas in English.

11.9 Summary of this chapter

The fidelity assessment indicates that EnglishFusion was successfully implemented in the participating schools. Teachers received effective training during formal sessions and weekly follow-ups, demonstrating the feasibility of infusing CT into the English curriculum at secondary schools in China.

The positive impact of EnglishFusion on CT skills is evident both in the classroom and beyond. However, some students and teachers remain sceptical about its effect on academic performance. This scepticism could be due to students' inability to transfer EnglishFusion content to their academic learning contexts, or a perception that EnglishFusion content is irrelevant to regular learning. Additionally, CT skills are not typically assessed in academic tests, which may contribute to the ineffectiveness. Furthermore, it may take longer to make the progress of academic attainment observable.

Practical challenges suggested by teachers and students, including limited teaching time, heavy workload, teaching to the test, poor English proficiency, and large class sizes, underscore the need for further exploration. Conditions for successfully implementing EnglishFusion are summarised to inform broader and improved educational practices.

Section V Discussion and conclusions

This section consists of two chapters. Chapter 12 presents a summary of the research findings, provides answers to the research questions, and responds to relevant literature. Given that this thesis includes both a systematic review and a randomised controlled trial, the limitations of each methodology are addressed. Chapter 13 offers implications and recommendations for educational practice, policy-making and future research.

Chapter 12 Discussion

This chapter summarises key research findings and discusses how these findings relate to previous literature. The summary is structured according to the research questions. Additionally, this chapter addresses the limitations of the thesis project, including both the systematic review and the randomised controlled trial.

12.1 Summary of findings

Findings are summarised based on the research questions. The first research question is answered through the systematic review, while the remaining questions are addressed by the main trial.

RQ1: What is the evidence on Chinese students' critical thinking compared with students of other nationalities?

The systematic review was conducted to examine the CT aptitude of Chinese students. Some studies analysed the CT skills of Chinese students exclusively, but these studies do not provide comparative insights. To determine whether Chinese students lack CT, it is necessary to compare their CT levels with those of students from other nationalities. Therefore, the review only considers studies that included a comparison group.

Fifteen studies were included in the review, measuring CT across three domains — CT skills, CT dispositions, and CT styles. Eight studies focused on Chinese students' CT skills, but their results were mixed. There is no evidence to support claims that Chinese students have higher or lower CT skills than students from other countries. The research in this area is currently inadequate. Six studies on CT dispositions suggest that Chinese students are less disposed to CT, which is not the same as being weak in CT skills. Only one study examined CT style, indicating that Chinese students prefer information seeking to engaging in CT styles.

All of the studies had some methodological weaknesses, including small sample sizes, high attrition or low response rates, the use of convenience sampling, and poor analytical processes. Most studies did not account for confounders, or establish group equivalence. These issues compromise the validity of the findings, necessitating caution in interpreting the results.

The overall conclusion is that we still do not have a definitive answer. There is no evidence that Chinese students have higher or lower CT skills or dispositions for CT. Additionally, there are no clear insights into the CT styles of Chinese students. More robust, larger-scale experimental studies are needed. Research in this field needs to improve.

RQ2: Can critical thinking skills be taught to Chinese secondary students who are not traditionally exposed to critical thinking?

Atkinson (1997) asserted that CT cannot be taught to individuals from non-Western cultures. Since Chinese secondary school students are not traditionally exposed to CT, it remains uncertain whether CT skills can be taught to them and whether there are any objections from teachers and students regarding CT instruction.

The cluster randomised controlled trial indicates that the infusion CT method was well received by both teachers and students in China. Although there were concerns about the link between CT skills and academic assessments, there was no resistance to the delivery and reception of CT lessons. Positive comments and feedback were provided by both teachers and students during interviews. The high degree of student engagement in EnglishFusion lessons also demonstrated the feasibility and popularity of the infusion approach. Moreover, teachers successfully implemented CT content in their English classes using various teaching resources, such as lesson plans, student handouts, and slides. The suggestion to formally infuse CT into the English curriculum and extend it to other educational levels indicates the appropriateness and potential of this intervention in China. This aligns with Lin's (2014) study, where CT was successfully infused into the English curriculum of Chinese high schools.

RQ3a: Does EnglishFusion improve Chinese secondary students' critical thinking skills?

The small positive effect size of +0.14 suggests that the infusion of CT into the English curriculum can improve Chinese secondary school students' CT skills. This aligns with previous studies demonstrating the effectiveness of the infusion method in teaching CT (e.g. Bağ & Gürsoy, 2021; Zohar & Tamir, 1993; Zohar et al., 1994). It also provides

evidence that CT skills can be enhanced with a three-month intervention (Niu, Behar-Horenstein, & Garvan, 2013). As CT skills are considered an overarching set of interdependent thinking skills in this study, the trial also shows the positive impact of the infusion method on skills such as argumentation, assumption, deduction, and inference.

The effectiveness of the approach can be attributed to the well implementation of the infusion teaching (Harn, Parisi, & Stoolmiller, 2013; O'Donnell, 2008). First, practical resources were provided to teachers, helping them adhere to CT teaching content and offering sufficient examples of CT instruction. This support made teaching CT easier, increasing teachers' enthusiasm for the CT instruction. Secondly, both teachers and students responded positively to the intervention, which likely increased students' receptiveness. Teachers benefited from the infusion method and recommended the formal incorporation of CT into regular education in China. Students expressed interest in the intervention and engaged highly in the CT lessons. Thirdly, experimental teachers generally prepared well for the infusion CT course, assisted by weekly follow-up training sessions. These sessions allowed teachers to reflect on class interactions and receive timely feedback, resulting in improved CT teaching over time. This contrasts with Mahmood's (2017) study, where inadequate preparation led to negative results.

However, EnglishFusion did not lead to an improvement in students' interpretation skills. This may be due to the possibility that interpretation skills were not sufficiently addressed in the classroom, as no specific lesson was dedicated to this area. Additionally, it is possible that interpretation skills require more time to develop, and their effects may not become immediately observable. A long-term impact evaluation might be necessary to accurately assess this.

RQ3b: Does EnglishFusion have a differential impact on the critical thinking skills of sub-groups of students (by age, birth sex, ethnicity, prior academic attainment, prior critical thinking skills, schools, parental involvement in children's education, and home background)?

The trial aims to explore whether the infusion CT approach is equally effective for students with different characteristics. Results of sub-group analyses show that

EnglishFusion is more effective for younger students, minority students, students from lower SES, higher academic achievers and higher critical thinkers.

However, these findings should be treated with caution due to the small sample sizes in some sub-groups. For example, only 39 students identified as minority, constituting just 2% of the overall participants. Individual changes in such small groups might exaggerate the final outcomes.

RQ4a: Does EnglishFusion improve Chinese secondary students' academic performance?

Given the emphasis on academic scores in Chinese secondary schools, the study investigated the impact of EnglishFusion on students' academic performance. The marginal negative effect size of -0.10 indicates that the intervention did not improve overall academic scores. Specifically, students who participated in EnglishFusion course did not show better progress in Chinese, Maths, and English compared to their counterparts.

This finding contrasts with Lin's (2014) study, which observed improvements in the English writing of 89 Chinese high school students after CT infusion teaching. Similarly, Hu et al. (2011) found that thinking teaching improved academic attainment in Chinese and Maths subjects among 116 Chinese primary school students. However, these results should be treated with caution due to several limitations. A major concern is the small number of students recruited from only one school, making it unclear whether the improvement in academic performance would be observed in a broader population. Additionally, both studies introduced some biases. In Lin's (2014) study, the assessment of English writing was jointly designed by the researcher and the participating teacher, potentially leading to the issue of teaching to the test since they knew what would be tested. In Hu et al.'s (2011) study, problematic randomisation could have caused a diffusion issue, as both experimental and control students were in the same class. Therefore, it is premature to conclusively assert a positive impact on academic attainment based on these studies.

The results of this trial should also be interpreted with caution. The Number Needed to Disturb (NNTD) of the academic scores suggests that the impact on Chinese and Maths attainment could be influenced by missing values. Additionally, the experimental group had much lower average scores in all academic subjects than the control group at the outset. This disparity is possibly due to class segregation, where students with similar academic performance were placed in the same class. Since this study conducted randomisation at the teacher level, segregation bias could have been introduced.

To examine the extent to which students' post-academic scores could be predicted by their background, prior academic scores, and the intervention, regression analysis was conducted. The results demonstrate that students' prior academic performance is the primary predictor of their post-academic performance. This finding is consistent with the study by Gorard, See, and Siddiqui (2014), which found that the treatment group membership had little effect on literacy attainment progress once pupils' demographics and prior attainment were considered.

EnglishFusion has a small positive impact on CT skills but not on academic attainment. Improving academic performance may require a longer intervention or stronger dose than enhancing CT skills. For instance, a meta-analysis indicates a larger association between CT and academic achievement when students are assessed one year later (Fong, Kim, Davis, Hoang, & Kim, 2017). Longer interventions or follow-up assessments may better evaluate the impact on academic achievement.

Another explanation is that CT skills are rarely included in Chinese academic assessments, which focus on curricular knowledge rather than thinking skills (Dong, 2015). Teachers in the trial believed that EnglishFusion course did not improve their students' English scores and expressed concerns about reduced time for textbook teaching. These concerns highlight that CT skills, or at least the infusion CT content, are not included in academic assessments. This aligns with the rote learning and knowledge-oriented assessments in China. Even in university CT courses, summative tests often check memorisation of thinking rules (Dong, 2015). Therefore, reforming academic assessments is necessary to cultivate CT skills effectively.

RQ4b: Does EnglishFusion have a differential impact on the academic attainment of sub-groups of students (by age, birth sex, ethnicity, prior academic attainment, prior critical thinking skills, schools, parental involvement in children's education, and home background)?

Given the emphasis on academic attainment in Chinese secondary school education, the trial considers it a secondary outcome and conducts a sub-group analysis to evaluate if EnglishFusion is equally effective for different students. The findings suggest that the intervention has a smaller negative effect on the academic attainment of younger students, boys and students with a lower parental involvement. EnglishFusion is more effective for students from lower SES background and those from Schools B, C and D.

These findings are tentative due to the small sample sizes in some sub-groups, such as minority students, and the initial disparity in academic attainment between the experimental and control groups. Although the imbalance was addressed using the progress score of academic attainment, the regression results indicate that students' post-academic attainment was predominantly predicted by their prior academic scores, and the intervention did not make any difference.

RQ5: Does training and teaching EnglishFusion alter teachers' critical awareness and attitudes towards teaching critical thinking?

The trial also measured changes in teachers' critical awareness and their attitudes toward CT teaching in the English curriculum. Teachers who used the infusion approach tended to place less credibility on external factors such as publication source, time, and authority. They were also more sceptical about standardised test data and large sample sizes. This means that EnglishFusion does not only improve students' CT skills, but also helpful in raising teachers' critical awareness. It should be noted that the conclusion is tentative as there are only 21 teachers in total. A larger sample size is needed to draw a more convincing conclusion.

Teachers who participated in the infusion CT training strongly believed that CT should be taught in schools. Although there was a minor decrease in their agreement with the statement that CT was relevant to the English curriculum, they still had a higher level of agreement than the control teachers. This marginal decrease may be due to the limited

CT content included in academic, particularly English, tests. This is echoed by the test-oriented perspective when teachers and students were asked to judge the usefulness of the intervention on their English language learning. Many teachers and students prioritised English language tests and felt that some EnglishFusion content might confuse students when answering close-ended reading questions.

12.2 Limitations

Limitations of the systematic review

As with any large-scale review, some relevant studies may have been missed. The key question is whether including these missed studies would have altered the results. Admittedly, limiting the review to English and Chinese language records published between 2000 and 2021 means that some potentially useful earlier studies may have been missed.

Additionally, the systematic review relies on existing literature. Despite efforts to include all levels of education, most research still focuses on higher education. To the best of our knowledge, no studies compare the CT performance of Chinese students with other learners at the primary, secondary, or high school levels. This suggests a research gap in this area. Notably, the majority of studies on this topic are conducted in the nursing discipline. It is not clear why this is so, and why comparisons of Chinese students' CT with other nationalities are not more widely studied in other disciplines.

Finally, despite frequent claims that Chinese students are deficient in CT (Song, 2014; Xu, 2021), surprisingly few studies have tested this assertion. CT involves multiple dimensions, but no single study has explored CT skills, dispositions, and styles simultaneously.

Limitations of the primary research

The trial also has several limitations. First, the randomisation was conducted at the teacher's level rather than the individual level. While this approach is pragmatic and often adopted in studies evaluating interventions (e.g. Bağ & Gürsoy, 2021; El Soufi, 2019), it does not eliminate the issue of class-level segregation. Some participating schools grouped students with similar academic levels into the same class. Thus, high-

achieving students are likely clustered in one class, while lower-scoring students form another. This cluster-level randomisation resulted in imbalanced prior academic scores between the experimental and control groups.

Secondly, the intervention lasted only three months. Behar-Horenstein and Niu (2011) argue that a more noticeable change in CT is observed when the intervention exceeds four months. Thus, the evaluation of the infusion CT approach in the English curriculum was based on short-term effects, leaving the long-term impact unknown. Additionally, the elapsed time before the post-test was relatively short, and many not have been long enough for any improvement in CT to translate into academic progress.

Thirdly, despite the large sample size, student dropout makes the evaluation of the intervention's impact on academic performance tentative. Although the dropout reasons were unrelated to the intervention, a conclusive claim about its impact on academic attainment cannot be made.

Furthermore, the class sizes in the trial were predominantly over 50 students, making it difficult for teachers to give adequate attention to each student and keep all students engaged. More individualised attention and support from teachers could potentially yield better results.

Another limitation is that the intervention was conducted in four rural schools. Students from these schools may have different characteristics from the general Chinese student population. For example, they usually do not start learning English until secondary school, unlike urban students who begin learning English in primary school or even kindergarten. Additionally, students in this study may have lower SES than their urban counterparts. Although including rural secondary schools enriches the current CT education research in China, which predominantly focuses on urban schools, generalising the findings from this study should be done cautiously.

Finally, there may be biases since I am both the developer of the intervention and the evaluator. This is a common issue in PhD intervention studies (El Soufi, 2019). I designed all the teaching resources, provided practical teacher training, and translated

the CT tests and student questionnaires. While I did not teach CT to the students directly, the evaluation process might still contain biases. Future research should involve independent evaluators to assess the effectiveness of an intervention.

Chapter 13 Implications and conclusions

This chapter discusses the implications and recommendations of both the systematic review and the randomised controlled trial (RCT). It considers teaching practice, policy-making and future research before drawing an overall conclusion.

13.1 Implications of the systematic review

The findings of this review provide no conclusive evidence that Chinese students are less capable of critical thinking (CT) compared to students of other nationalities. With no evidence either, the notion that Chinese students exhibit weaker CT skills should be set aside for now. A lack of critical awareness and scepticism is not unique to Chinese students (See, 2016). For instance, Arum and Roksa (2011) found that over two thousand American university graduates struggle to distinguish facts from opinions, construct clear arguments, and objectively assess conflicting reports. Similarly, Sampson and Walker (2012) identified that undergraduates from a US college had difficulty in providing strong evidence in scientific writing to support their arguments. Additionally, Zhao and Liao (2024) discovered that both Chinese students and their Western counterparts tended to prioritise superficial characteristics of academic reports, such as titles, authors, and publication dates, when evaluating them. These examples suggest that students, regardless of nationality, may not possess sufficient CT skills. Misrepresenting Chinese students as lacking criticality could lead to inappropriate educational interventions.

There is, however, tentative evidence that there are differences in the CT skills, dispositions, and styles of Chinese students compared to students of other nationalities, but overall, the evidence is weak. The lack of high-quality studies suggests that this area is under-researched, possibly due to the uncritical acceptance of the belief that Chinese students have lower CT skills. This stereotype is perpetuated by some academics in Western universities, who portray Chinese students as passive and uncritical (Atkinson, 1997; Cortazzi & Jin, 1997; Zhang, 2017). Such perceptions can be damaging, as they may prevent students from developing their CT skills if they internalise this stereotype. Academics who accept this view may inadvertently reinforce it by treating Chinese students as passive learners. Developing curricula and pedagogical approaches in Western universities to support Chinese students' CT skills

might be a misguided effort if their perceived passivity and reluctance to voice disagreement are mistakenly interpreted as a lack of criticality.

Given the weak evidence in this area, further research is needed to improve our understanding. To generate high-quality findings, future studies should consider the following:

- Employing randomised controlled designs or equivalent methodologies that control for both observable and unobservable factors, such as demographic background, socio-economic status, qualifications, and prior academic attainment, to ensure group equivalence.
- Using larger samples across a range of schools in various contexts or geographical regions, enabling the findings to be generalised to the wider population.
- Selecting standardised and independent instruments, with careful consideration of language and testing environments to avoid teaching to the test.
- Expanding research beyond higher education and the nursing discipline to include mainstream schools.

13.2 Implications and recommendations of the trial

The results from the RCT demonstrate that infusing CT into the curriculum is not only feasible but also beneficial for developing CT skills among Chinese secondary school students. This finding aligns with previous research (Bağ & Gürsoy, 2021; Lin, 2014; Zohar & Tamir, 1993). While the advancement of CT skills has been outlined in the English curriculum standards for secondary schools (MoE, 2022b), this represents merely the initial stage of implementation. Continued efforts are required. The findings from the primary research have implications for practice, policy and research.

Implications for teaching practice

While EnglishFusion shows promise for enhancing CT skills, its potential negative impact on academic achievement presents a challenge that requires careful consideration in educational practice. Lesson observations found that teachers continue to rely on traditional methods of instruction, where the teacher delivers information and

students passively accept it. There are limited opportunities for open debate or questioning. Students initially felt uncomfortable asking questions, and teachers often responded quickly with answers. In situations where there are no clear right or wrong answers, requiring students to argue and explain their responses, teachers felt lost and lacked control.

While the trial has demonstrated the feasibility of infusing CT into the regular curriculum, it is clear that teachers need training to effectively deliver these lessons. Teachers need to familiarise themselves with new pedagogical approaches. The explicit teaching of CT requires teachers to act as facilitators and instigators of thoughtful discussions, encouraging debates and questioning assumptions. This has implications for teacher development.

Classroom observations indicate that teachers are not adequately prepared to teach CT due to their own educational backgrounds, which have typically followed a traditional, teacher-centred approach that prioritises rote memorisation. To teach CT explicitly in Chinese classrooms, comprehensive professional development is necessary to help teachers adopt and effectively implement CT strategies (Paul & Elder, 2010). Currently, most teachers in China participate in CT training for administrative reasons (Dong, 2015), with few attending out of personal interest. It is recommended that CT training be incorporated into both teachers' professional development and initial teacher training to advance CT education in China.

Implications for policy-making

The findings of the study also have implications for educational policies, particularly in relation to a comprehensive overhaul of the exam system and revisions to school curricula.

The primary research findings indicate that the intervention is particularly beneficial for Chinese secondary school students, especially younger ones, suggesting they are well-positioned to gain the full benefits of CT instruction. This is consistent with other studies (e.g. Fung, 2017; Ku et al., 2014; Wang et al., 2017).

However, the study found no impact on academic attainment. Several possible reasons may explain this. One reason could be that the transference of CT skills to academic achievement takes time. Improving academic performance is challenging, and time is required for this transference to occur. This finding suggests that CT skills may need to be introduced earlier in schools to allow for this transfer effect. Early exposure to CT helps in developing essential cognitive skills, building a strong foundation for both personal and academic growth, and fostering independent thinking (Kuhn, 1999; O'Reilly, Devitt, & Hayes, 2022; Pollarolo, Størksen, Skarstein, & Kucirkova, 2023).

Another reason could be that learning CT skills, which are complex and require a new way of thinking, while simultaneously learning content in secondary school, may negatively impact students' academic achievement. This suggests the need for a gradual integration of the infusion method, allowing students time to adapt to this pedagogical approach. Gradual learning of CT may help ensure that both CT and subject content are adequately covered and practised.

A third reason is that current academic assessments primarily test knowledge recall rather than thinking skills. The discrepancy between CT skills and academic attainment implies that traditional academic metrics may not fully capture or reward the cognitive growth that CT promotes. This suggests the need to revamp assessment methods to better evaluate students' cognitive development rather than their memorisation skills.

There is currently a conflict between developing CT skills and meeting the demands of the current educational system, where academic achievement is often narrowly defined by exam performance. The heavy reliance on high-stakes exams, such as the Gaokao (China's university entrance examination), which emphasises content learning over problem-solving, has made it difficult for schools to implement CT in regular lessons. It is often the case that what is not tested will not be taught. Some students in the trial admitted to not paying attention to EnglishFusion course because it was not assessed, and teachers focused more on preparing students for academic tests.

To encourage the teaching of CT in the classroom, national exams, such as the Gaokao, need to be revamped to include more open-ended questions that require original thought

and problem-solving. CT skills should be assessed through integration with curricular content rather than through the memorisation of CT theories.

In conjunction with revamping the examination system, current textbooks will need to be revised. Some existing English textbook content is outdated and does not align with recent curriculum standards. Therefore, revisions to textbooks may be necessary to better balance and value both CT and traditional academic performance.

The findings from the RCT suggest that students from lower socioeconomic status (SES) backgrounds showed substantial growth in both CT skills and academic attainment. The infusion method of CT teaching appears to be particularly beneficial for this group of students, aiding them in learning and transferring CT content that they may not have been exposed to elsewhere. This may contrast with the views of some teachers who believe the intervention might have a greater impact on urban students, who tend to be more proficient in English and possess broader background knowledge. As the intervention could help reduce educational disparities by providing lower SES students with tools to improve their CT and academic outcomes, it is recommended that the intervention be introduced to schools in rural and more deprived areas.

Recommendations for future research

The trial has shown no beneficial impact of the infusion of CT on students' academic achievement. It is possible that the effects on academic attainment take longer to manifest. Future research should investigate the long-term impact of the intervention. Alternatively, future studies could explore why the development of CT skills does not appear to translate into improved academic attainment, perhaps through a pilot trial, which also considers changes in assessment methods.

While the trial indicates that the training and teaching of the infusion approach could raise teachers' critical awareness, the sample size was small. Future studies should explore the intervention's impact on teachers.

This evaluation is limited to the English curriculum in secondary schools. It remains uncertain whether the feasibility and promising effects would persist in other academic

disciplines and across different geographical and cultural contexts. Future research should explore the integration of CT into subjects such as Maths and Science, and extend the investigation to primary and high schools. A replication study could also be considered to test its impact on other parts of China.

The extent to which English is used as the medium of instruction depends largely on students' English language proficiency. The schools involved in this trial were in rural areas, where most students did not begin learning English until secondary school. It remains to be seen whether increased use of English in EnglishFusion lessons would enhance students' English skills. This is a critical area for future research.

13.3 Conclusions

This thesis provides a comprehensive evaluation of the common assumption that Chinese students lack a critical mind and assesses the effectiveness of the infusion CT approach on Chinese secondary students' CT skills and academic attainment. The findings from the systematic review reveal a lack of robust evidence to definitively conclude whether Chinese students possess higher or lower CT skills compared to students of other nationalities. There is a tentative indication of differences in CT dispositions and styles, with Chinese students potentially showing less disposition towards CT and a greater inclination towards an information-seeking style. Methodological weaknesses in the reviewed studies, such as small sample sizes and lack of control for confounding factors, highlight the need for more rigorous and larger-scale experimental studies in this area.

The stereotypical perception of Chinese students as passive learners lacking criticality may be detrimental, as it could reinforce these stereotypes rather than encourage the development of CT skills. Moreover, the review indicates that the perceived lack of criticality is not unique to Chinese students, as similar concerns exist in other educational systems globally.

Regarding the impact of the intervention, the results indicate a small positive effect on CT skills, particularly in domains such as argument, assumption, deduction, and inference. However, the findings also suggest that the infusion approach does not

impact overall academic performance, as measured by Chinese, Maths, and English scores. The discrepancy between the enhancement of CT skills and the stagnation in academic performance raises questions about the alignment of infusion CT teaching with the current test-oriented educational system in China. Additionally, sub-group analyses demonstrate that certain groups, such as younger students, minority students, and those with lower SES, showed greater improvement in CT skills. However, the limited sample size in some sub-groups warrants cautious interpretation of these results. Furthermore, the trial highlighted changes in teachers' critical awareness and attitudes towards CT teaching, with experimental teachers demonstrating increased critical awareness, as well as greater confidence and engagement in implementing CT in their English classes.

These findings have implications for teaching practice, policy, and future research. Recommendations for teaching practice include incorporating CT instruction into teacher professional development, enabling educators to transition to a student-centred approach where students are encouraged to question assumptions, engage in debate, and explore alternative perspectives. Policy recommendations should focus on introducing CT to students at an early age, gradually integrating CT into the curriculum, revamping assessment methods, reforming textbooks to accommodate CT, and expanding the infusion method to rural schools. Further research is needed to examine the long-term effects of the intervention, explore its impact on teachers, and investigate its implementation in the broader educational landscape. By addressing these evidence-based recommendations, educators, policymakers and researchers can better support the development of CT skills among Chinese students.

References

Studies included in the structured review and synthesis are marked with *

- Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. *Review of educational research*, 85(2), 275–314.
- Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R., & Zhang, D. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of Educational Research*, 78(4), 1102–1134.
- Ahn, E., & Kang, H. (2018). Introduction to systematic review and meta-analysis. *Korean journal of anesthesiology*, 71(2), 103–112.
- Akins, J. L., Lamm, A. J., Telg, R., Abrams, K., Meyers, C., & Raulerson, B. (2019). Seeking and Engaging: Case Study Integration to Enhance Critical Thinking About Agricultural Issues. *Journal of Agricultural Education*, 60(3), 97–108.
- Al-Ghadouni, A. M. (2021). Instructional approaches to critical thinking: an overview of reviews. *Revista Argentina de Clínica Psicológica*, 30(1), 240–246.
- Andrews, R. (2007). Argumentation, critical thinking and the postgraduate dissertation. *Educational Review*, 59(1), 1–18. <https://doi.org/10.1080/00131910600796777>
- Arum, R., & Roksa, J. (2011). *Academically adrift: Limited learning on college campuses*. Chicago: University of Chicago Press.
- Atkinson, D. (1997). A Critical Approach to Critical Thinking in TESOL. *TESOL Quarterly*, 31(1), 71–94. <https://doi.org/10.2307/3587975>
- Atkinson, D. (1998). Comments on Dwight Atkinson's "A Critical Approach to Critical Thinking in TESOL": A Case for Critical Thinking in the English Language Classroom. The Author Responds. *TESOL Quarterly*, 32(1), 133–137.
- Badger, J. (2019). A Case Study of Chinese Students' and IEP Faculty Perceptions of a Creativity and Critical Thinking Course. *Higher Education Studies*, 9(3), 34–44.
- Bağ, H. K., & Gürsoy, E. (2021). The effect of critical thinking embedded English course design to the improvement of critical thinking skills of secondary school learners. *Thinking Skills and Creativity*, 41, 1–13.

- Bailin, S. (2002). Critical thinking and science education. *Science & Education*, 11(4), 361–375.
- Bailin, S., Case, R., Coombs, J. R., & Daniels, L. B. (1999). Conceptualizing critical thinking. *Journal of Curriculum Studies*, 31(3), 285–302.
- Baker, M., Lu, P., & Lamm, A. J. (2021). Assessing the dimensional validity and reliability of the University of Florida Critical Thinking Inventory (UFCTI) in Chinese: a confirmatory factor analysis. *Journal of International Agricultural and Extension Education*, 28(3), 41–56.
- Barnett, R., & Davies, M. (2015). *The Palgrave handbook of critical thinking in higher education*. Basingstoke: Palgrave Macmillan.
- Behar-Horenstein, L. S., & Niu, L. (2011). Teaching critical thinking skills in higher education: A review of the literature. *Journal of College Teaching & Learning*, 8(2), 25–42.
- Bennett, Z., Faltin, L., & Wright, M. (2003). Critical thinking and international postgraduate students. *Discourse: Learning and Teaching in Philosophical and Religious Studies*, 3(1), 63–94.
- Bensley, D. A., & Spero, R. A. (2014). Improving critical thinking skills and metacognitive monitoring through direct infusion. *Thinking Skills and Creativity*, 12, 55–68.
- Bernard, R., Zhang, D., Abrami, P., Sicol, F., Borokhovski, E., & Surkes, M. (2008). Exploring the structure of the Watson–Glaser Critical Thinking Appraisal: One scale or many subscales? *Thinking Skills and Creativity*, 3(1), 15–22.
- Black, B. (2007). Critical Thinking—a tangible construct. *Research Matters: A Cambridge Assessment Publication*, 2, 2–4.
- Boland, A., Cherry, M. G., & Dickson, R. (2017). *Doing a systematic review: a student's guide* (2nd ed.). Los Angeles: Sage.
- Bondy, K. N., Koenigseder, L. A., Ishee, J. H., & Williams, B. G. (2001). Psychometric properties of the California critical thinking tests. *Journal of Nursing Measurement*, 9, 309–328.
- Bourdieu, P. (1986). The forms of capital. In J. Richardson (Ed.), *Handbook of Theory and Research for the Sociology of Education* (pp. 241–258). New York: Greenwood Press.

- Boynton, P. M., & Greenhalgh, T. (2004). Selecting, designing, and developing your questionnaire. *BMJ*, 328(7451), 1312–1315.
- Bridgeman, B., & Moran, R. (1996). Success in college for students with discrepancies between performance on multiple choice and essay tests, *Journal of Educational Psychology*, 88, 333–340.
- Bugge, C. (2024). What is a process evaluation when used alongside a randomised controlled trial?. *Evidence-Based Nursing*, 27(2), 45–47.
- Bycio, P., & Allen, J. S. (2009). The California Critical Thinking Skills Test and Business School Performance. *American Journal of Business Education (AJBE)*, 2(8), 1–8.
- Byrne, M. (1994). *Learning to be critical*. Newcastle: Material and Resources Centre for Enterprising Teaching.
- Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological bulletin*, 119(2), 197–253.
- Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review*, 48(3), 378–399.
- Chan, N. M., Ho, I. T., & Ku, K. Y. (2011). Epistemic beliefs and critical thinking of Chinese students. *Learning and Individual Differences*, 21(1), 67–77.
- Chan, S., & Wong, F. (1999). Development of basic nursing education in China and Hong Kong. *Journal of Advanced Nursing*, 29(6), 1300–1307.
- Chen, L. (2017). Understanding critical thinking in Chinese sociocultural contexts: A case study in a Chinese college. *Thinking Skills and Creativity*, 24, 140–151.
- Chen, M., & Shi, N. (2017). Brief review in research advance of critical thinking in China and other countries. *International Journal of Education, Culture and Society*, 2(1), 13–19.
- Cheng, K. M. (2010). Shanghai and Hong Kong: Two distinct examples of education reform in China. *Organisation for economic co-operation and development, strong performers and successful performers in education: Lessons from PISA for the United States*, 83–115.
- Cheng, Y. C., Huang, L. C., Yang, C. H., & Chang, H. C. (2020). Experiential learning program to strengthen self-reflection and critical thinking in freshmen nursing

- students during COVID-19: A quasi-experimental study. *International journal of environmental research and public health*, 17(15), 5442.
- Cheung, A., & Slavin, R. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, 45(5), 283–292.
- Choy, S. C., & Cheah, P. K. (2009). Teacher perceptions of critical thinking among students and its influence on higher education. *International Journal of teaching and learning in Higher Education*, 20(2), 198–206.
- Clifford, J. S., Boufal, M. M., & Kurtz, J. E. (2004). Personality traits and critical thinking skills in college students: Empirical tests of a two-factor theory. *Assessment*, 11(2), 169–176.
- Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society open science*, 1(3), 1–16.
- Cortazzi, M., & Jin, L. (1997). Communication for learning across cultures. In D. McNamara & R. Harris (Eds.), *Overseas students in higher education: Issues in teaching and learning* (pp. 76–90). London: Routledge.
- Cortazzi, M., Pilcher, N., & Jin, L. (2011). Language choices and ‘blind shadows’: investigating interviews with Chinese participants. *Qualitative Research*, 11(5), 505–535.
- Costa, P. T., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological assessment*, 4(1), 5–13.
- Critical Appraisal Skills Programme. (2021). *CASP checklists*. Retrieved from <https://casp-uk.net/casp-tools-checklists/>
- Cui, R., & Teo, P. (2023). Thinking through talk: Using dialogue to develop students’ critical thinking. *Teaching and Teacher Education*, 125, 104068. <https://doi.org/10.1016/j.tate.2023.104068>.
- Davies, M. (2013). Critical thinking and the disciplines reconsidered. *Higher Education Research & Development*, 32(4), 529–544.
- Deal, K. D., & Pittman, J. (2009). Examining predictors of social work students’ critical thinking skills. *Advances in Social Work*, 10(1), 87–102.
- Dello-Iacovo, B. (2009). Curriculum reform and ‘quality education’ in China: An overview. *International journal of educational development*, 29(3), 241–249.

- *Dennett, S. K. (2014). *A Study to Compare the Critical Thinking Dispositions between Chinese and American College Students*. Retrieved from <https://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED557684&site=ehost-live>
- Dennett, S. K., & DeDonno, M. A. (2021). A comparison between Chinese and American male and female college students' critical thinking dispositions. *International Journal of Chinese Education*, 10(3), 22125868211046966.
- Dewey, J. (1909). *How we think*. New York: D.C. Health and Company.
- Dewey, J. (1933). *How we think: A restatement of the relation of reflective thinking to the educative process*. New York: D.C. Health and Company.
- Dhakal, K. R., Watson Todd, R., & Jaturapitakkul, N. (2023). Unpacking the nature of critical thinking for educational purposes. *Educational Research and Evaluation*, 28(4-6), 130–151.
- Din, M. (2020). Evaluating university students' critical thinking ability as reflected in their critical reading skill: A study at bachelor level in Pakistan. *Thinking Skills and Creativity*, 35, 100627.
- Dong, Y. (2010). *Principles and Methods of Critical Thinking: Toward the New Knowledge and Action*. Beijing: China Higher Education Press.
- Dong, Y. (2015). Critical Thinking Education with Chinese Characteristics. In Davies, M., Barnett, R. (Eds.), *The Palgrave Handbook of Critical Thinking in Higher Education* (pp. 351–368). Palgrave Macmillan, New York. https://doi.org/10.1057/9781137378057_22
- Dong, Y. (2017). Teaching and assessing critical thinking in second language writing: An infusion approach. *Chinese Journal of Applied Linguistics*, 40(4), 431–451. <https://doi.org/10.1515/cjal-2017-0025>
- *Dong, Y.X., Li, K., & Liu, F. (2010). The critical thinking skills of college English students: Assessment and cultivation. *Computer-assisted Foreign Language Educational in China*, 135, 33–38.
- Doyal, L. (2003). Sex and gender: the challenges for epidemiologists. *International Journal of Health Services*, 33(3), 569–579.
- Dumville, J. C., Torgerson, D. J., & Hewitt, C. E. (2006). Reporting attrition in randomised controlled trials. *Bmj*, 332(7547), 969–971.

- Durišić, M., & Bunijevac, M. (2017). Parental involvement as a important factor for successful education. *Center for Educational Policy Studies Journal*, 7(3), 137–153.
- Durkin, K. (2008). The adaptation of East Asian masters students to western norms of critical thinking and argumentation in the UK. *Intercultural Education*, 19(1), 15–27.
- Durkin, K. (2011). Adapting to western norms of critical argumentation and debate. In J. Lixian & M. Cortazzi (Eds.), *Researching Chinese Learners: Skills, Perceptions and Intellectual Adaptations* (pp. 274–291). New York: Palgrave Macmillan.
- Dwyer, C. P., Hogan, M. J., & Stewart, I. (2014). An integrated critical thinking framework for the 21st century. *Thinking Skills and Creativity*, 12, 43–52. <https://doi.org/10.1016/j.tsc.2013.12.004>.
- El Soufi, N. (2019). *Evaluating the impact of instruction in critical thinking on the critical thinking skills of English language learners in higher education* (Doctoral dissertation, Durham University).
- El Soufi, N., & See, B. H. (2019). Does explicit teaching of critical thinking improve critical thinking skills of English language learners in higher education? A critical review of causal evidence. *Studies in educational evaluation*, 60, 140–162.
- Elder, L., & Paul, R. (2020). *Critical thinking: Tools for taking charge of your learning and your life*. Foundation for Critical Thinking.
- Emir, S. (2009). Education faculty students' critical thinking disposition according to achedemic achievement. *Procedia-Social and Behavioral Sciences*, 1(1), 2466–2469.
- Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. *Educational Leadership*, 43(2), 44–48.
- Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In J. B. Baron & R. J. Sternberg (Eds.), *Teaching thinking skills: Theory and practice* (pp.9–26). New York: W.H. Freeman and Company.
- Ennis, R. H. (1989). Critical thinking and subject specificity: Clarification and needed research. *Educational researcher*, 18(3), 4–10.

- Ennis, R. H. (2011). The nature of critical thinking: An outline of critical thinking dispositions and abilities. *University of Illinois*, 2(4), 1–8.
- Ennis, R. H. (2015). Critical Thinking: A Streamlined Conception. In M. Davies & R. Barnett (Eds.), *The Palgrave Handbook of Critical Thinking in Higher Education* (pp.31–47). New York: Palgrave Macmillan.
- Ennis, R. H., Millman, J., & Tomko, T. N. (2005a). *Cornell Critical Thinking Tests* (Fifth Edition). Seaside, CA: The Critical Thinking Company.
- Ennis, R. H., Millman, J., & Tomko, T. N. (2005b). *Cornell Critical Thinking Tests Levels X and Z Administration Manual*. Seaside, CA: Critical Thinking Company.
- Erasmus, A., Holman, B., & Ioannidis, J. P. (2022). Data-dredging bias. *BMJ Evidence-Based Medicine*, 27(4), 209–211.
- Evans, J.S.B.T. (2005). Deductive Reasoning. In Holyoak, K.J. & Morrison, R.G. (Eds.), *The Cambridge Handbook of Thinking and Reasoning* (pp.169–184). United States of America: Cambridge University Press.
- Facione, P. A. (1990). Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction. The Delphi report: Research findings and recommendations. In *ERIC Doc. No. ED315-423*. ERIC Washington.
- Facione, P. A., & Facione, N. C. (1992). *The California Critical Thinking Disposition Inventory*. Millbrae, CA: California Academic Press.
- Facione, P. A., Facione, N. C., & Giancarlo, C. A. (2000). The Disposition Towards Critical Thinking: Its Character, Measurement and Relationship to Critical Thinking Skill. *Informal Logic*, 20(1), 61–84.
- Facione, P. A., Sánchez, C. A., Facione, N. C., & Gainen, J. (1995). The disposition toward critical thinking. *The Journal of General Education*, 44(1), 1–25.
- Fakunle, L., Allison, P., & Fordyce, K. (2016). Chinese postgraduate students' perspectives on developing critical thinking on a UK education masters. *Journal of Curriculum and Teaching*, 5(1), 27–38.
- Fan, K. (2022). [Review of the book *Infusing critical thinking into your course: a concrete practical approach*, by Linda B. Nilson]. *Educational Research and Evaluation*, 1–3. DOI: 10.1080/13803611.2022.2143132

- Feeley, N., Cossette, S., Côté, J., Héon, M., Stremler, R., Martorella, G., & Purden, M. (2009). The importance of piloting an RCT intervention. *Canadian Journal of Nursing Research Archive*, 84–99.
- Feng, R. C., Chen, M. J., Chen, M. C., & Pai, Y. C. (2010). Critical thinking competence and disposition of clinical nurses in a medical center. *Journal of Nursing Research*, 18(2), 77–87.
- Fife-Schaw, C. (2001). Questionnaire design. In G. M. Breakwell, S. Hammond, & C. Fife-Schaw (Eds.), *Research methods in psychology* (pp. 158–176). London: Sage.
- Fisher, A. (2011). *Critical thinking: An introduction*. Cambridge university press.
- Floyd, C. B. (2011). Critical thinking in a second language. *Higher Education Research and Development*, 30(3), 289–302.
- Fong, C. J., Kim, Y., Davis, C. W., Hoang, T., & Kim, Y. W. (2017). A meta-analysis on critical thinking and community college student achievement. *Thinking Skills and Creativity*, 26, 71–83.
- Fox, H. (1994). *Listening to the World: Cultural issues in academic writing*. Urbana, IL: National Council of Teachers of English.
- Fung, D. (2014). Promoting critical thinking through effective group work: A teaching intervention for Hong Kong primary school students. *International Journal of Educational Research*, 66, 45–62.
- Fung, D. (2017). The pedagogical impacts on students' development of critical thinking dispositions: Experience from Hong Kong secondary schools. *Thinking Skills and Creativity*, 26, 128–139.
- Fung, D., & Howe, C. (2014). Group work and the learning of critical thinking in the Hong Kong secondary liberal studies curriculum. *Cambridge Journal of Education*, 44(2), 245–270.
- Gay, K. D., Terry, B., & Lamm, A. J. (2015). Identifying Critical Thinking Styles to Enhance Volunteer Development. *The Journal of Extension*, 53(6), Article 28. <https://doi.org/10.34068/joe.53.06.28>
- Gorard, S. (2013). *Research design: creating robust approaches for the social sciences*. London: Sage.
- Gorard, S. (2020). Handling missing data in numeric analyses. *International Journal of Social Research Methodology*, 23(6), 651–660.

- Gorard, S. (2021). *How to Make Sense of Statistics: Everything You Need to Know about Using Numbers in Social Science*. London: Sage.
- Gorard, S., & Gorard, J. (2016). What to do instead of significance testing? Calculating the ‘number of counterfactual cases needed to disturb a finding’. *International Journal of Social Research Methodology*, 19(4), 481–490. <https://doi.org/10.1080/13645579.2015.1091235>
- Gorard, S., & See, B. H. (2013). *Overcoming disadvantage in education*. Routledge.
- Gorard, S., See, B. H., & Siddiqui, N. (2014). *Switch-On Reading: Evaluation Report and Executive Summary*. Education Endowment Foundation.
- Gorard, S., See, B. H., & Siddiqui, N. (2017). *The trials of evidence-based education: The promises, opportunities and problems of trials in education*. Taylor & Francis.
- Gorham, L. M., Lamm, A. J., & Rumble, J. N. (2014). The critical target audience: Communicating water conservation behaviors to critical thinking styles. *Journal of Applied Communications*, 98(4), 42–55.
- Greenholtz, J. (2003). Socratic teachers and Confucian learners: Examining the benefits and pitfalls of a year abroad. *Language and Intercultural Communication*, 3(2), 122–130.
- Gu, Z., & Liu, Z. (2006). *Critical thinking text*. Beijing: Peking University Press.
- Guo, L., & O'Sullivan, M. (2012). From Laoshi to partners in learning: Pedagogic conversations across cultures in an international classroom. *Canadian Journal of Education*, 35(3), 164–179.
- Hagstrom, C., Kendall, S., & Cunningham, H. (2015). Googling for grey: using Google and Duckduckgo to find grey literature. In *Abstracts of the 23rd Cochrane Colloquium. Cochrane database systematic reviews supplements* (pp. 1–327).
- Halpern, D. (1998). Teaching critical thinking for transfer across domains: Dispositions, skills, structure training, and metacognitive monitoring. *American Psychologist*, 53(4), 449–455.
- Halpern, D. (1999). Teaching for critical thinking: Helping college students develop the skills and dispositions of a critical thinker. *New directions for teaching and learning*, 1999(80), 69–74.

- Halpern, D. (2005). *Halpern Critical Thinking Assessment: Background and scoring standards*. Unpublished manuscript. Claremont, CA: Claremont McKenna College.
- Hammersley M. (2020). Reflections on the Methodological Approach of Systematic Reviews. In: Zawacki-Richter O., Kerres M., Bedenlier S., Bond M., Buntins K. (Eds.) *Systematic Reviews in Educational Research* (pp.23–39). Springer VS, Wiesbaden.
- Hare, W. (1999). Critical thinking as an aim of education. In R. Marples (Ed.), *The aims of education* (pp. 85–99). London: Routledge.
- Haritania, H., Febrianib, Y., Yulianac, T. P., & Arviana, E. (2019). The correlation of undergraduate course research experience and critical thinking skills. *International Journal of Innovation, Creativity and Change*, 5(6), 336–347.
- Hariton, E., & Locascio, J. J. (2018). Randomised controlled trials—the gold standard for effectiveness research. *BJOG: an international journal of obstetrics and gynaecology*, 125(13), 1716.
- Harlen, W., & Crick, R.D. (2004). Opportunities and challenges of using systematic reviews of research for evidence-based policy in education. *Evaluation & Research in Education*, 18(1-2), 54–71.
- Harn, B., Parisi, D., & Stoolmiller, M. (2013). Balancing fidelity with flexibility and fit: What do we really know about fidelity of implementation in schools?. *Exceptional Children*, 79(2), 181–193.
- Harpe, S. E. (2015). How to analyze Likert and other rating scale data. *Currents in pharmacy teaching and learning*, 7(6), 836–850.
- Hassan, K. E., & Madhum, G. (2007). Validating the Watson Glaser critical thinking appraisal. *Higher Education*, 54, 361–383.
- Heng, T. T. (2018). Different is not deficient: Contradicting stereotypes of Chinese international students in US higher education. *Studies in higher education*, 43(1), 22–36.
- Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2021). *Cochrane handbook for systematic reviews of interventions*. Retrieved from www.training.cochrane.org/handbook

- Horn, S., & Veermans, K. (2019). Critical thinking efficacy and transfer skills defend against ‘fake news’ at an international school in Finland. *Journal of Research in International Education*, 18(1), 23–41.
- *Hu, L. Q., Adelopo, I., & Last, K. (2020). Understanding Students’ Critical Thinking Ability: A Comparative Case of Chinese and British Undergraduates. *New Educational Review*, 61, 133–143.
- Hu, W., Adey, P., Jia, X., Liu, J., Zhang, L., Li, J., & Dong, X. (2011). Effects of a ‘Learn to Think’ intervention programme on primary school students. *British journal of educational psychology*, 81(4), 531–557.
- Huang, R. (2008). Critical thinking: Discussion from Chinese postgraduate international students and their lecturers. *Hospitality, Leisure, Sport and Tourism Education*, 4(23), 1–12.
- Huang, Y. (2019). Establishing Critical Thinking Course for High School Students in China: A Literature Review in Pedagogy Field. *Proceedings of the International Conference on Education*, 5(1), 59–66.
- Huazhong University of Science and Technology (2011). *The Critical Thinking course has been well received and Qiming College intends to introduce it to the whole country*. Retrieved from <https://news.hust.edu.cn/info/1004/21986.htm>
- Hwang, G. J., Huang, H., Wang, R. X., & Zhu, L. L. (2021). Effects of a concept mapping-based problem-posing approach on students’ learning achievements and critical thinking tendency: An application in Classical Chinese learning contexts. *British Journal of Educational Technology*, 52(1), 374–493.
- Independent. (2006). *University students: They Can’t Write, Spell or Present an Argument*. Retrieved from <http://www.independent.co.uk/news/education/higher/university-students-they-cant-write-spell-or-present-an-argument-479536.html>
- Indra, V. (2019). Critical Thinking Disposition in Asian and Non-Asian Countries: A Review. *International Journal of Nursing Education and Research*, 7(2), 279–282.
- Ioannidis J. P. (2005). Why most published research findings are false. *PLoS medicine*, 2(8), e124. <https://doi.org/10.1371/journal.pmed.0020124>

- Ip, W. Y., Lee, D. T., Lee, I. F., Chau, J. P., Wootton, Y. S., & Chang, A. M. (2000). Disposition towards critical thinking: a study of Chinese undergraduate nursing students. *Journal of Advanced Nursing*, 32(1), 84–90.
- Jacob, S. M. (2012). Analyzing critical thinking skills using online discussion forums and CCTST. *Procedia-Social and Behavioral Sciences*, 31, 805–809.
- Jiang, J. (2013). Critical thinking in general education in China. *International Journal of Chinese Education*, 2(1), 108–134.
- Joanna Briggs Institute. (2017). *Checklist for Systematic Reviews and Research Syntheses*. Retrieved from https://jbi.global/sites/default/files/2019-05/JBI_Critical_Appraisal-Checklist_for_Systematic_Reviews2017_0.pdf
- Johnson, R. H. (1992). The problem of defining critical thinking. In S.P. Norris (Ed.), *The generalizability of critical thinking: Multiple perspectives on an educational ideal* (pp. 38–53). New York: Teacher College Press.
- Kamin, C., O’Sullivan, P., & Deterding, R. (2002, April). Does project L.I.V.E. case modality impact critical thinking in PBL groups? Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED464921)
- Kirkbride, P. S., Tang, S. F., & Westwood, R. I. (1991). Chinese conflict preferences and negotiating behaviour: Cultural and psychological influences. *Organization studies*, 12(3), 365–386.
- Klassen, T. P., Jadad, A. R., & Moher, D. (1998). Guides for reading and interpreting systematic reviews: I. Getting started. *Archives of pediatrics & adolescent medicine*, 152(7), 700–704.
- Koretz, D. (2006). *Measuring Up: What educational testing really tells us*. Cambridge, MA: Harvard University Press.
- Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into practice*, 41(4), 212–218.
- Ku, K. Y. (2009). Assessing students’ critical thinking performance: Urging for measurements using multi-response format. *Thinking skills and creativity*, 4(1), 70–76.
- *Ku, K. Y. L., Chan, N.-M., Lun, V. M.-C., Haplern, D. F., Marin-Burkhart, L., Hau, K. T., & Ho, I. T. (2006, April). Chinese and United States undergraduates' critical thinking skills: Academic and dispositional predictors. Paper presented at the

- 2006 Annual Meeting of the American Educational Research Association: Education Research in the Public Interest, San Francisco, California.
- Ku, K. Y., Ho, I. T., Hau, K. T., & Lai, E. C. (2014). Integrating direct and inquiry-based instruction in the teaching of critical thinking: an intervention study. *Instructional Science*, 42, 251–269.
- Kuhn, D. (1999). A developmental model of critical thinking. *Educational Researcher*, 28(2), 16–25. <https://doi.org/10.3102/0013189X028002016>.
- Lai, E. R. (2011). Critical thinking: A literature review. *Pearson's Research Reports*, 6, 1–49.
- Lamm, A. J. (2015). Integrating critical thinking into extension programming #3: Critical thinking style. Florida Cooperative Extension Service Electronic Data Information Source AEC546. <https://edis.ifas.ufl.edu/wc208>
- Lamm, A. J., & Irani, T. (2011). *UFCTI manual*. Gainesville, FL: University of Florida.
- Lawson, T. J., Jordan-Fleming, M. K., & Bodle, J. H. (2015). Measuring psychological critical thinking. *Teach. Psychol.* 42, 248–253. 10.1177/0098628315587624
- *Lee, H. Y., Kim, Y., Kang, H., Fan, X. Z., Ling, M., Yuan, Q. H., & Lee, J. (2011). An international comparison of Korean and Chinese nursing students with nursing curricula and educational outcomes. *Nurse Education Today*, 31(5), 450–455. <https://doi.org/10.1016/j.nedt.2010.09.002>
- Lee, H.-S., Liu, O. L., & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education*, 24(2), 115–136.
- Lee, K. S., & Carrasquillo, A. (2006). Korean college students in United States: Perceptions of professors and students. *College Student Journal*, 40(2), 442–457.
- Leung, S. O. (2011). A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point Likert scales. *Journal of social service research*, 37(4), 412–421.
- Li, G., Chen, W., & Duanmu, J. L. (2010). Determinants of international students' academic performance: A comparison between Chinese and other international students. *Journal of studies in international education*, 14(4), 389–405.
- Li, X. (2017). Exploration and Reflection in Dialogue and Development—The exploration and practice of cultivating pupils' critical thinking in the Affiliated

- Primary School of Huazhong University of Science and Technology. *New Curriculum Research*, (6), 31–33.
- Li, Y. (2013). First year ESL students developing critical thinking: Challenging the stereotypes. *Journal of Education and Training Studies*, 1(2), 186–196.
- Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. *Journal of clinical epidemiology*, 62(10), e1–e34.
- Lin, Y. (2014). *Infusion of critical thinking into L2 classes: a case study in a Chinese high school*. (Doctoral dissertation, Newcastle University).
- Lipman, M. (1998). Teaching students to think reasonably: Some findings of the Philosophy for Children program. *The Clearing House*, 71(5), 277–280.
- Lipman, M. (2003). *Thinking in education* (2nd ed.). Cambridge: Cambridge University Press.
- *Liu, H. (2013). *Assessment of Students' Critical Thinking Skills in College English Program 2012* international conference on education reform and management innovation, 1, 436–441.
- Liu, O. L., Frankel, L., & Roohr, K. C. (2014). *Assessing critical thinking in higher education: Current state and directions for next-generation assessment (ETS RR 14–10)*. Princeton, NJ: Educational Testing Service.
- Liu, S., & Onwuegbuzie, A. J. (2012). Chinese teachers' work stress and their turnover intention. *International journal of educational research*, 53, 160–170.
- Liu, X., & Lu, K. (2008). Student performance and family socioeconomic status: Results from a survey of compulsory education in Western China. *Chinese Education & Society*, 41(5), 70–83.
- Liu, Y., & Pásztor, A. (2022). Design and validate the employer-employee-supported critical thinking disposition inventory (2ES-CTDI) for undergraduates. *Thinking Skills and Creativity*, 46, 101169.
- *Loyalka, P., Liu, O. L., Li, G., Kardanova, E., Chirikov, I., Hu, S., Yu, N., Ma, L., Guo, F., Beteille, T., Tognatta, N., Gu, L., Ling, G., Federiakin, D., Wang, H., Khanna, S., Bhuradia, A., Shi, Z., & Li, Y. (2021). Skill levels and gains in university STEM education in China, India, Russia and the United States. *Nature Human*

Behaviour, 5(7), 892–904. <https://doi.org/http://dx.doi.org/10.1038/s41562-021-01062-3>

- *Lu, P., Burris, S., Baker, M., Meyers, C., & Cummins, G. (2021). Cultural Differences in Critical Thinking Style: A Comparison of US and Chinese Undergraduate Agricultural Students. *Journal of International Agricultural and Extension Education*, 28(4), 49–62.
- Lu, S., & Singh, M. (2017). Debating the capabilities of “Chinese students” for thinking critically in anglophone universities. *Education Sciences*, 7, 1–16.
- Lucas, K. J. (2019). Chinese graduate student understandings and struggles with critical thinking: A narrative-case study. *International Journal for the Scholarship of Teaching and Learning*, 13(1), 1–7.
- *Lun, V. M. C., Fischer, R., & Ward, C. (2010). Exploring cultural differences in critical thinking: Is it about my thinking style or the language I speak?. *Learning and Individual differences*, 20(6), 604–616.
- Mahmood, S. (2017). *Testing the effectiveness of a critical thinking skills intervention for initial teacher education students in Pakistan* (Doctoral dissertation, University of Southampton).
- Manalo, E., & Sheppard, C. (2016). How might language affect critical thinking performance?. *Thinking skills and Creativity*, 21, 41–49.
- Marin, L., & Halpern, D. F. (2011). Pedagogy for developing critical thinking in adolescents: Explicit instruction produces greatest gains. *Thinking Skills and Creativity*, 6, 1–13.
- *McBride, R.E., Xiang, P., Wittenberg, D. & Shen, J. (2002). An analysis of preservice teachers’ dispositions toward critical thinking: A cross-cultural perspective. *Asia–Pacific Journal of Teacher Education*, 30(2), 131–140.
- McCoy, C. E. (2017). Understanding the intention-to-treat principle in randomized controlled trials. *Western Journal of Emergency Medicine*, 18(6), 1075–1078.
- McCutcheon, L. E., Hanson, E., Apperson, J. M., & Wynn, V. (1992). Relationships among critical thinking skills, academic achievement, and misconceptions about psychology. *Psychological reports*, 71(2), 635–639.
- McGuire, J. M. (2007). Why has the critical thinking movement not come to Korea?. *Asia Pacific Education Review*, 8, 224–232.
- McPeck, J. E. (1981). *Critical thinking and education*. Oxford: Robertson.

- Ministry of Education. (2005). *Chronology of Quality Education*. Retrieved from http://www.moe.gov.cn/jyb_xwfb/xw_zt/moe_357/s3579/moe_1081/tnull_12374.html
- Ministry of Education. (2022a). *Compulsory Education Curriculum Plan (2022 edition)*. Retrieved from <http://www.moe.gov.cn/srcsite/A26/s8001/202204/W020220420582343217634.pdf>
- Ministry of Education. (2022b). *English curriculum standards for compulsory education*. Retrieved from <http://www.moe.gov.cn/srcsite/A26/s8001/202204/W020220420582349487953.pdf>
- Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., Elbourne, D., Egger, M., & Altman, D. G. (2010). CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *Journal of Clinical Epidemiology*, *63*(8), 1–28.
- Moore, T. (2013). Critical thinking: Seven definitions in search of a concept. *Studies in Higher Education*, *38*(4), 506–522.
- Moosavi, L. (2021). The myth of academic tolerance: the stigmatisation of East Asian students in Western higher education. *Asian Ethnicity*, 1–20.
- Newman, M., & Gough, D. (2020). Systematic reviews in educational research: Methodology, perspectives and application. In: Zawacki-Richter O., Kerres M., Bedenlier S., Bond M., Buntins K. (Eds.) *Systematic Reviews in Educational Research* (pp.3–22). Springer VS, Wiesbaden.
- Ng, S. Y., Cheung, K., & Cheng, H. L. (2022). Critical thinking cognitive skills and their associated factors in Chinese community college students in Hong Kong. *Sustainability*, *14*(3), 1127.
- Nilson, L. B. (2021). *Infusing Critical Thinking Into Your Course: A Concrete, Practical Approach*. Stylus Publishing, LLC.
- Niu, L., Behar-Horenstein, L. S., & Garvan, C. W. (2013). Do instructional interventions influence college students' critical thinking skills? A meta-analysis. *Educational research review*, *9*, 114–128.
- Norris, S. P. (1989). Can we test validly for critical thinking?. *Educational researcher*, *18*(9), 21–26.

- Norris, S. P. (2003). The meaning of critical thinking test performance: The effects of abilities and dispositions on scores. *Critical thinking and reasoning: Current research, theory and practice*, 315–329.
- O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of educational research*, 78(1), 33–84.
- O'Reilly, C., Devitt, A., & Hayes, N. (2022). Critical thinking in the preschool classroom-A systematic literature review. *Thinking skills and creativity*, 46, 101110.
- O'Sullivan, M., & Guo, L. (2010). Critical thinking and Chinese international students: An East-West dialogue. *Journal of Contemporary Issues in Education*, 5, 53–73.
- Oakley, A., Strange, V., Bonell, C., Allen, E., & Stephenson, J. (2006). Process evaluation in randomised controlled trials of complex interventions. *Bmj*, 332(7538), 413–416.
- Oakley, A., Strange, V., Toroyan, T., Wiggins, M., Roberts, I., & Stephenson, J. (2003). Using random allocation to evaluate social interventions: three recent UK examples. *The Annals of the American Academy of Political and Social Science*, 589(1), 170–189.
- OECD. (2013). *Education at a Glance 2013: Highlights*. OECD Publishing, Paris, https://doi.org/10.1787/eag_highlights-2013-en.
- OECD. (2014). *PISA 2012 results: Creative problem solving: Students' skills in tackling real-life problems (Volume V)*. PISA: OECD Publishing.
- OECD. (2018a). *Future of education and skills: Skills for 2020*. Paris: OECD. Retrieved from [https://www.oecd.org/education/2030/E2030%20Position%20Paper%20\(05.04.2018\).pdf](https://www.oecd.org/education/2030/E2030%20Position%20Paper%20(05.04.2018).pdf)
- OECD. (2018b). *OECD Programme for International Student Assessment 2018*. Retrieved from https://www.oecd.org/pisa/data/2018database/CY7_201710_QST_MS_P_AQ_NoNotes_final.pdf
- OECD. (2021). *Study on Social and Emotional Skills (SSES) Main Study Student Questionnaire (younger Cohort)*. Retrieved

- from [https://webfs.oecd.org/sses/Test%20Form%202. _Student%20\(Younger%20cohort\).pdf](https://webfs.oecd.org/sses/Test%20Form%202. _Student%20(Younger%20cohort).pdf)
- Ofsted. (2024). *School inspection handbook*. Retrieved from <https://www.gov.uk/government/publications/school-inspection-handbook-eif/school-inspection-handbook-for-september-2023>
- Oppenheim, A. (2000) *Questionnaire design, interviewing and attitude measurement*. London: Continuum.
- Orhan, A. (2022). California critical thinking disposition inventory: reliability generalization meta-analysis. *Journal of Psychoeducational Assessment, 40*(2), 202–220.
- Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hrobjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., et al., (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ 372*, n71. <https://doi.org/10.1136/>
- Paine, L., & Fang, Y.P. (2006). Reform as hybrid model of teaching and teacher development in China. *International Journal of Education Research, 45*, 279–289.
- *Park, J. H., Niu, W. H., Cheng, L., & Allen, H. (2021). Fostering Creativity and Critical Thinking in College: A Cross-Cultural Investigation. *Frontiers in Psychology, 12*. <https://doi.org/10.3389/fpsyg.2021.760351>
- Paton, M. (2005). Is critical analysis foreign to Chinese students? In E. Manalo & G. Wong-Toi (Eds.), *Communication skills in university education: The international dimension* (pp. 1–11). Auckland: Pearson Education.
- Paul, R. & Elder, L. (2012). *Critical Thinking: Tools for Taking Charge of Your Learning and Your Life*. 3rd Edition. Boston: Pearson Education, Inc.
- Paul, R. W., & Binker, A. J. A. (1990). *Critical thinking: What every person needs to survive in a rapidly changing world*. Center for Critical Thinking and Moral Critique, Sonoma State University, Rohnert Park, CA 94928.
- Paul, R. W., & Elder, L. (2006). Critical thinking: The nature of critical and creative thought. *Journal of Developmental Education, 30*(2), 34–35.

- Paul, R., & Elder, L. (2010). *The Miniature Guide to Critical Thinking Concepts and Tools*. Dillon Beach: Foundation for Critical Thinking Press.
- Paul, R., Elder, L., & Bartell, T. (1997). *A brief history of the idea of critical thinking*. Retrieved from <https://www.criticalthinking.org/pages/a-brief-history-of-the-idea-of-critical-thinking/408>
- Paulhus, D. L., & Vazire, S. (2007). The self-report method. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality* (pp. 224–239). London, England: Guilford.
- Paulhus, D.L. (1991). Measurement and control of response bias. In J. P. Robinson, P.R. Shaver, & L.S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.
- People's Republic of China. (2010). *Outline of China's national plan for medium and long-term education reform and development (2010-2020)*. Beijing: People's Publishing House.
- *Petrini, M. A., & Kawashima, A. (2003). Comparison of Critical Thinking Skills of Nurses in Japan, China and Samoa. *Bulletin of the Graduate Schools Yamaguchi Prefectural University*, 4, 11–32.
- Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Oxford: Blackwell.
- Ping, W. (2010). A case study of an in-class silent postgraduate Chinese student in London Metropolitan University: A journey of learning. *TESOL Journal*, 2(1), 207–214.
- Pintrich, P. R., Smith, D., Garcia, T., & McKeachie, W. (1991). *A manual for the use of the motivated strategies for learning questionnaire (MSLQ)*. Ann Arbor, MI: The University of Michigan.
- PISA. (2018). *Released FT and MS Reading Literacy Items*. Retrieved from https://www.oecd.org/pisa/test/PISA2018_Released_REA_Items_121120_19.pdf
- Pollarolo, E., Størksen, I., Skarstein, T. H., & Kucirkova, N. (2023). Children's critical thinking skills: Perceptions of Norwegian early childhood educators. *European Early Childhood Education Research Journal*, 31(2), 259–271.

- Puig, B., Blanco-Anaya, P., Bargiela, I. M., & Crujeiras-Pérez, B. (2019). A systematic review on critical thinking intervention studies in higher education across professional fields. *Studies in Higher Education, 44*(5), 860–869.
- QAA. (2002). *Quality Assurance Agency Subject Benchmark Statements: Biosciences*. Cheltenham: Quality Assurance Agency for higher Education.
- Qasserras, M., & Qasserras, L. (2023). Critical thinking: A western guise or a thinking, cultural, and pedagogical fatigue. *World Journal of Advanced Research and Reviews, 19* (01), 273–283. <https://doi.org/10.30574/wjarr.2023.19.1.1354>
- Quinn, S., Hogan, M., Dwyer, C., Finn, P., & Fogarty, E. (2020). Development and validation of the student-educator negotiated critical thinking dispositions scale (SENCTDS). *Thinking Skills and Creativity, 38*, Article 100710.
- Ramanathan, V., & Kaplan, R. B. (1996). Audience and voice in current L1 composition texts: Some implications for ESL student writers. *Journal of Second Language Writing, 5*(1), 21–34.
- Rear, D. (2019). One size fits all? The limitations of standardised assessment in critical thinking. *Assessment & Evaluation in Higher Education, 44*(5), 664–675.
- Reboot Foundation. (2021). *The State of Critical Thinking 2021*. Retrieved from https://reboot-foundation.org/wp-content/uploads/_docs/Critical_Thinking_Survey_Report_2021.pdf
- Rimiene, V. (2002). Assessing and developing students' critical thinking. *Psychology Learning & Teaching, 2*(1), 17–22.
- Ro, J. (2023). Critical thinking in the national curriculum and teacher education in South Korea: a missing link. *Teachers and Teaching, 1–18*.
- Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in methods and practices in psychological science, 1*(1), 27–42.
- Ryan, J., Kang, C., Mitchell, I., & Erickson, G. (2009). China's basic education reform: An account of an international collaborative research and development project. *Asia Pacific Journal of Education, 29*(4), 427–441.
- Salsali, M., Tajvidi, M., & Ghiyasvandian, S. (2013). Critical thinking dispositions of nursing students in Asian and non-Asian countries: A literature review. *Global Journal of Health Science, 5*(6), 172–178.

- Sampson, V., & Walker, J. P. (2012). Argument-driven inquiry as a way to help undergraduate students write to learn by learning to write in chemistry. *International Journal of Science Education*, *34*(10), 1443–1485.
- Schraw, G., Crippen, K. J., & Hartley, K. (2006). Promoting self-regulation in science education: Metacognition as part of a broader perspective on learning. *Research in Science Education*, *36* (1–2), 111–139.
- See, B. H. (2018). Evaluating the evidence in evidence-based policy and practice: Examples from systematic reviews of literature. *Research in Education*, *102*(1), 37–61.
- See, B.H. (2016). An investigation into the teaching and learning of argumentation in first-year undergraduate courses: A pilot study. *British Journal of Education, Society and Behavioural Science*, *18* 4, 1–25.
- Semerci, N. (2006). The effect of problem-based learning on the critical thinking of students in the intellectual and ethical development unit. *Social Behavior and Personality*, *34*(9), 1127–1136.
- Şendağ, S., & Odabaşı, H. F. (2009). Effects of an online problem based learning course on content knowledge acquisition and critical thinking skills. *Computers & education*, *53*(1), 132–141.
- Sheikh, K., & Mattingly, S. (1981). Investigating non-response bias in mail surveys. *Journal of Epidemiology & Community Health*, *35*(4), 293–296.
- Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019). How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual review of psychology*, *70*, 747–770.
- Sievers, K. (2001). How do you know that...? *Critical Reasoning Handbook*. Adelaide, Australia: Philosophy Department, Flinders University.
- Slavin, R. & Neitzel, A. (2020). *In meta-analyses, weak inclusion standards lead to misleading conclusions. Here's proof.* Retrieved from <https://robertslavinsblog.wordpress.com/category/published-vs-unpublished/>
- Slavin, R. (2017). *Reviewing Social and Emotional Learning for ESSA: MOOSES, not Parrots.* Retrieved from <https://robertslavinsblog.wordpress.com/2017/05/25/reviewing-social-and-emotional-learning-for-essa-moooses-not-parrots/>

- Slavin, R. (2020). *Even Magic Johnson sometimes had bad games: Why research reviews should not be limited to published studies*. Retrieved from <https://robertslavinsblog.wordpress.com/2020/02/27/even-magic-johnson-sometimes-had-bad-games-why-research-reviews-should-not-be-limited-to-published-studies/>
- Snyder, L. G., & Snyder, M. J. (2008). Teaching critical thinking and problem solving skills. *The Journal of Research in Business Education*, 50(2), 90–99.
- Snyder, S. J., Edwards, L. C., & Sanders, A. L. (2019). An empirical model for infusing critical thinking into higher education. *Journal on Excellence in College Teaching*, 30(1), 127–156.
- Solon, T. (2007). Generic critical thinking infusion and course content learning in introductory psychology. *Journal of Instructional Psychology*, 34(2), 95-109.
- Song, F., Hooper, L., & Loke, Y. (2013). Publication bias: what is it? How do we measure it? How do we avoid it?. *Open Access Journal of Clinical Trials*, 2013(5), 71–81.
- Song, X. (2014). Changing social relations in higher education: the first-year international students and the ‘Chinese learner’ in Australia. In H. Brook, D. Fergie, M. Maeorg, & D. Mitchell (Eds.), *Universities in Transition: Foregrounding Social Contexts of Knowledge in the First Year Experience* (pp. 127–156). University of Adelaide Press.
- Sosu, E. M. (2013). The development and psychometric validation of a Critical Thinking Disposition Scale. *Thinking skills and creativity*, 9, 107–119.
- Spence, P. (2012). *Parental involvement in the lives of college students: Impact on student independence, self-direction, and critical thinking*. Doctoral dissertation. Chicago: Loyola University.
- Stapleton, P. (2011). A survey of attitudes towards critical thinking among Hong Kong secondary school teachers: Implications for policy change. *Thinking Skills and Creativity*, 6(1), 14–23.
- Stein, M. L., Berends, M., Fuchs, D., McMaster, K., Sáenz, L., Yen, L., Fuchs, L.S., & Compton, D. L. (2008). Scaling up an early reading program: Relationships among teacher support, fidelity of implementation, and student performance across different sites and years. *Educational evaluation and policy analysis*, 30(4), 368–388.

- Sternberg, R. J. (1986). *Critical Thinking: Its Nature, Measurement, and Improvement*. Washington, DC: National Institute for Education.
- Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the P value is not enough. *Journal of graduate medical education*, 4(3), 279–282.
- Suri, H. (2020). Ethical considerations of conducting systematic reviews in educational research. In: Zawacki-Richter O., Kerres M., Bedenlier S., Bond M., Buntins K. (Eds.) *Systematic Reviews in Educational Research* (pp.41–54). Springer VS, Wiesbaden.
- Tan, C. (2020). Conceptions and practices of critical thinking in Chinese schools: An example from Shanghai. *Educational Studies*, 56(4), 331–346.
- Tarran, B. (2019). Is this the end of “statistical significance”? *Significance*, 16(2), 4–5.
- Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical education*, 42(2), 198–206.
- Taube, K. T. (1997). Critical thinking ability and disposition as factors of performance on a written critical thinking test. *Journal of General Education*, 46, 129-164.
- Tawfik, G. M., Dila, K. A. S., Mohamed, M. Y. F., Tam, D. N. H., Kien, N. D., Ahmed, A. M., & Huy, N. T. (2019). A step by step guide for conducting a systematic review and meta-analysis with simulation data. *Tropical medicine and health*, 47(1), 1–9.
- Ten Dam, G., & Volman, M. (2004). Critical thinking as a citizenship competence: teaching strategies. *Learning and Instruction*, 14(4), 359–379.
- Thabane, L., Mbuagbaw, L., Zhang, S., Samaan, Z., Marcucci, M., Ye, C., ... & Goldsmith, C. H. (2013). A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. *BMC medical research methodology*, 13(1), 1–12.
- Tian, J., & Low, G. D. (2011). Critical thinking and Chinese university students: A review of the evidence. *Language, Culture and Curriculum*, 24(1), 61–76. <https://doi.org/10.1080/07908318.2010.546400>
- Tilbury, C., Osmond, J. & Scott, T. (2010). Teaching critical thinking in social work education: A literature review. *Advances in Social Work and Welfare Education*, 11, 31–50.

- Tiruneh, D. T., Verburch, A., & Elen, J. (2014). Effectiveness of critical thinking instruction in higher education: A systematic review of intervention studies. *Higher Education Studies*, 4(1), 1–17.
- Tiruneh, D. T.; De Cock, M. & Elen, J. (2018). Designing Learning Environments for Critical Thinking: Examining Effective Instructional Approaches. *International Journal of Science and Mathematics Education*, 16, 1065–1089.
- *Tiwari, A., Avery, A., & Lai, P. (2003). Critical thinking disposition of Hong Kong Chinese and Australian nursing students. *Journal of Advanced Nursing*, 44(3), 298–307.
- Tiwari, A., Lai, P., So, M., & Yuen, K. (2006). A comparison of the effects of problem-based learning and lecturing on the development of students' critical thinking. *Medical education*, 40(6), 547–554.
- Toomey, E., Hardeman, W., Hankonen, N., Byrne, M., McSharry, J., Matvienko-Sikar, K., & Lorencatto, F. (2020). Focusing on fidelity: narrative review and recommendations for improving intervention fidelity within trials of health behaviour change interventions. *Health Psychology and Behavioral Medicine*, 8(1), 132–151.
- Torgerson, C. (2003). *Systematic reviews*. London: Continuum.
- Torgerson, C.J. & Torgerson, D.J. (2001). The need for randomised controlled trials in educational research. *British Journal of Educational Studies*, 49(3), 316–328.
- Toy, B. Y., & Ok, A. (2012). Incorporating critical thinking in the pedagogical content of a teacher education programme: does it make a difference?. *European Journal of Teacher Education*, 35(1), 39–56.
- Turner, Y. (2006). Students from mainland China and critical thinking in postgraduate business and management degrees: Teasing out tensions of culture, style and substance. *International Journal of Management Education*, 5(1), 3–11.
- Van Gelder, T. (2005). Teaching critical thinking: Some lessons from cognitive science. *College Teaching*, 53(1), 41–48.
- Ventura, M., Lai, E., & DiCerbo, K. (2017). *Skills for Today: What We Know about Teaching and Assessing Critical Thinking*. London: Pearson.
- Walsh, C. M., & Hardy, R. C. (1997). Factor structure stability of the California critical thinking disposition inventory across sex and various students' majors. *Perceptual Motor Skills*, 85(3_suppl), 1211–1228.

- Wang, D., & Jia, Q. (2023). Twenty Years of Research Development on Teachers' Critical Thinking: Current Status and Future Implications—A Bibliometric Analysis of Research Articles Collected in WOS. *Thinking Skills and Creativity*, 101252.
- Wang, H. H., Chen, H. T., Lin, H. S., Huang, Y. N., & Hong, Z. R. (2017). Longitudinal study of a cooperation-driven, socio-scientific issue intervention on promoting students' critical thinking and self-regulation in learning science. *International Journal of Science Education*, 39(15), 2002–2026.
- Wang, S., & Seepho, S. (2017). Facilitating Chinese EFL Learners' Critical Thinking Skills: The Contributions of Teaching Strategies. *SAGE Open*, (1), 1–9.
- Watkins, D. A. & Biggs, J. B. (Eds.). (1996). *The Chinese learner: cultural, psychological, and contextual influences*. Hong Kong/Melbourne: CERC & ACER.
- Watson, G., & Glaser, E. M. (2012). *Watson-Glaser critical thinking appraisal (3rd ed.): User's guide and technical manual and user's guide*. London: Pearson Education Limited.
- Willingham, D. (2008). Critical thinking: why is it so hard to teach? *Arts Education Policy Review*, 109(4), 21–32.
- Willingham, D. (2019). *How to teach critical thinking*. Education Future Frontiers.
- Wu, H. (2011). Critical thinking: study of the semantics. *Journal of Yanan University (Social Science)*, 33(1), 5–17.
- Wu, H., & Leung, S. O. (2017). Can Likert scales be treated as interval scales? — A simulation study. *Journal of social service research*, 43(4), 527–532.
- Xu, C. L. (2021). Portraying the 'Chinese international students': a review of English-language and Chinese-language literature on Chinese international students (2015-2020). *Asia Pacific Education Review*, 23, 1–17. <https://doi.org/10.1007/s12564-021-09731-8>
- Xu, F. (2017). *Research proposal of critical thinking training of ordinary high school students*. Retrieved from http://www.njzhzx.net/_upload/article/82/fe/bbd722b649b7b6d83e4889b3504d/c4f8e230-c475-4e97-b23e-0cf48aceea7f.pdf

- Xu, F. (2019). The human nature perspective of critical thinking and its educational enlightenment. *Jiangsu Social Sciences*, 6, 146–154. doi:10.13858/j.cnki.cn32-1312/c.20191121.022.
- Yan, C. (2012). ‘We can only change in a small way’: A study of secondary English teachers’ implementation of curriculum reform in China. *Journal of Educational Change*, 13, 431–447.
- Yan, C. (2015). ‘We can’t change much unless the exams change’: Teachers’ dilemmas in the curriculum reform in China. *Improving Schools*, 18(1), 5–19.
- Yanchar, S. C., & Slife, B. D. (2004). Teaching critical thinking by examining assumptions. *Teaching of Psychology*, 31(2), 85–90.
- *Yeh, M. L., & Chen, H. H. (2003). Comparison affective dispositions toward critical thinking across Chinese and American baccalaureate nursing students. *Journal of Nursing Research*, 11(1), 39–46.
- Yoon, J. (2004). *Development of an instrument for the measurement of critical thinking disposition: In nursing*. (Doctoral dissertation, The Catholic University of Korea).
- Yuan, H., Kunaviktikul, W., Klunklin, A., & Williams, B. A. (2008). Improvement of nursing students’ critical thinking skills through problem-based learning in the People's Republic of China: A quasi-experimental study. *Nursing & Health Sciences*, 10(1), 70–76.
- Zawacki-Richter, O., Kerres, M., Bedenlier, S., Bond, M., & Buntins, K. (2020). *Systematic Reviews in Educational Research*. Wiesbaden: Springer.
- Zhai, J.Y. (2015). Critical thinking curriculum, why claim that Chinese students really need it?. *Critical and Creative Thinking Education Newsletter*, 24, 24–27.
- Zhang, C., Fan, H., Xia, J., Guo, H., Jiang, X., & Yan, Y. (2017). The effects of reflective training on the disposition of critical thinking for nursing students in China: A controlled trial. *Asian Nursing Research*, 11(3), 194–200.
- Zhang, H., & Lambert, V. (2008). Critical thinking dispositions and learning styles of baccalaureate nursing students from China. *Nursing & Health Sciences*, 10(3), 175–181. <https://doi.org/10.1111/j.1442-2018.2008.00393.x>
- Zhang, H., Yuan, R., & He, X. (2020). Investigating university EFL teachers’ perceptions of critical thinking and its teaching: Voices from China. *The Asia-Pacific Education Researcher*, 29, 483–493.

- *Zhang, Q., & Zhang, J. (2013). Instructors' positive emotions: Effects on student engagement and critical thinking in US and Chinese classrooms. *Communication Education*, 62(4), 395–411.
- Zhang, T. (2017). Why do Chinese postgraduates struggle with critical thinking? Some clues from the higher education curriculum in China. *Journal of Further and Higher Education*, 41(6), 857–871. doi:10.1080/0309877X.2016.1206857
- Zhao, P., & Liao, X. (2024). Advancing to the academics: How did first-year Chinese undergraduates evaluate academic literature?. *Thinking Skills and Creativity*, 101600. <https://doi.org/10.1016/j.tsc.2024.101600>
- Zheng, K. (2017). 2017 National advanced training course for critical thinking teachers. *Critical and Creative Thinking Education Newsletter*, 37, 31–32.
- Zhong, W., & Cheng, M. (2021). Developing critical thinking: experiences of Chinese International Students in a Post-1992 University in England. *Chinese Education & Society*, 54(3-4), 95–106.
- Zohar, A., & Tamir, P. (1993). Incorporating critical thinking into a regular high school biology curriculum. *School Science and Mathematics*, 93(3), 136–140.
- Zohar, A., Weinberger, Y., & Tamir, P. (1994). The effect of the biology critical thinking project on the development of critical thinking. *Journal of Research in Science Teaching*, 31(2), 183–196.
- Zou, M., & Lee, I. (2023). Learning to teach critical thinking: testimonies of three EFL teachers in China. *Asia pacific journal of education*, 43(3), 867–881.
- Zulkpli, Z., Abdullah, A. H., Kohar, U. H. A., & Ibrahim, N. H. (2017). A review research on infusion approach in teaching thinking: advantages and impacts. *Man In India*, 97(12), 289–298.

Appendices

Appendix A. Search syntax and results in databases

Databases	Search syntax	Numbers of records
Applied Social Sciences Index & Abstracts (ASSIA)	ab ("critical thinking" OR "think critically" OR "critical reasoning" OR "thinking skill*") AND ab (China OR Chinese) AND ab (student* OR learner* OR pupil*)	33
EBSCO host <ul style="list-style-type: none"> • Open dissertations • British Education Index • Education Abstracts • ERIC • APA PsycArticles • APA PsycInfo 	AB ("critical thinking" OR "think critically" OR "critical reasoning" OR "thinking skill*") AND AB (China OR Chinese) AND AB (student* OR learner* OR pupil*)	280
ProQuest: <ul style="list-style-type: none"> • Dissertations & Theses Global • ProQuest Social Sciences Premium 	ab ("critical thinking" OR "think critically" OR "critical reasoning" OR "thinking skill*") AND ab (China OR Chinese) AND ab (student* OR learner* OR pupil*)	497
Sage Journals	[[Abstract "critical thinking"] OR [Abstract "think critically"] OR [Abstract "critical reasoning"] OR [Abstract "thinking skill*"]] AND [[Abstract China] OR [Abstract Chinese]] AND [[Abstract student*] OR [Abstract learner*] OR [Abstract pupil*]]	20
Scopus	(TITLE-ABS-KEY ("critical thinking" OR "think critically" OR "critical reasoning" OR "thinking skill*") AND TITLE-ABS-KEY (China OR Chinese) AND TITLE-ABS-KEY (student* OR learner* OR pupil*) AND PUBYEAR > 1999 AND PUBYEAR < 2022	373
Web of Science	ab= ("critical thinking" OR "think critically" OR "critical reasoning" OR "thinking skill*") AND ab= (China OR Chinese) AND ab= (student* OR learner* OR pupil*)	257
Wiley online library	""critical thinking" OR "think critically" OR "critical reasoning" OR "thinking skill*" in Abstract and "China OR Chinese" in Abstract and "student* OR learner* OR pupil*" in Abstract	21
In total	1481	

Appendix B. Data extraction tables

Table 1. Studies on critical thinking skills (n=8)

Author(s) & date	Research design	Sample & level of education	Measuring instrument(s)	Finding(s) & result(s)	Limitation(s)	Rating
Loyalka et al. (2021)	A cross-sectional, cohort, comparative, and descriptive study	5,102 freshmen and 4,145 junior Chinese students, 8,232 freshmen and 9,223 third-year Indian students, 2,607 freshmen and 2,096 third-year Russian students, and 973 undergraduate U.S. students Sampling strategies in institutions: simple random sampling in China; stratified national random sampling in India and Russia; non-random sampling in the U.S. Sampling strategies within the sample institutions: random sampling in China, India and Russia; non-random sampling in the U.S. Undergraduate	Critical Thinking Exam, part of the HEIghten® suite of assessments from Educational Testing Service (ETS) Translated to native languages in China, India and Russia	The freshmen and second-year Chinese students show similar critical thinking skills levels as their American counterparts, whereas their Indian and Russian peers are far lower. Fourth-year Chinese university students demonstrate higher scores in critical thinking skills than Indian students, similar to Russian students, but much lower than the U.S. students in the fourth year. Minimal gains in critical thinking skills are exhibited in the first two years in Chinese, Indian and Russian students. Significant decrease in this aspect is evidenced in Chinese, Indian and Russian students during the last two years. On the contrary, American students show an increase in critical thinking skills during the final half of the university life.	Only focus on two disciplines (computer science and electrical engineering) Not necessarily generalize to other contexts	3*

				A mixed result		
Hu, Adelopo, & Last (2020)	A cross-sectional study	50 British students and 50 Chinese students Not clear about sampling strategy Final-year undergraduate	Watson-Glaser Critical Thinking Appraisal questionnaire (WGCTA) Form S Modified: Content reduced to 20 questions in 5 sections (4 questions per section) Translated to a Chinese version	Chinese students' inference skill score is higher than that of their counterparts (55% vs 51%). However, scores of assumption, arguments and interpretation skills of Chinese students are lower than those of the English cohort, with 51% vs 72%, 41% vs 50%, and 58% vs 63% respectively. The deduction skill scores between the two groups are similar, with 63% of English students and 62.5% of Chinese students. Overall, Chinese students' critical thinking skills are poorer than that of British students. A mixed result	Small scale study, restricted in only one UK university A short duration of research time Not a full WGCTA test	2*
Ku et al. (2006)	A correlational, cross-sectional study	142 Chinese students (43 males, 99 females) and 153 U.S. students (30 males, 121 females, 2 with missing gender information) Not clarify the sampling strategy Undergraduate	Halpern Critical Thinking Assessment Using Everyday Situations (HCTAES) Translated to Chinese language	Chinese students (mean 119.20, SD 14.33) gained higher scores than U.S. students (mean 108.92, SD 18.11) in terms of the critical thinking test. Higher CT skills	The cross-sectional design	2*
Dong, Li, & Liu (2010)	A descriptive and comparative study	25 Chinese undergraduates (8 females, 17 males)	The California Critical Thinking Skills Test (CCTST)-2000 designed	Chinese students' comprehensive critical thinking skills scores (mean 19.20, SD	Small sample size, hard to be representative	1*

		Stratified random sampling Final-year undergraduate	by California Assessment Center (CAC) Translated to Chinese language	4.32) are higher than those of American students (mean 16.80, SD 5.06). Chinese students demonstrate a lower level in analysis (mean 3.52, SD 1.33 vs mean 4.44, SD 1.41) and induction (mean 9.32, SD 2.32 vs mean 9.53, SD 2.82), while higher in inference (mean 10.32, SD 2.40 vs mean 7.85, SD 2.69), evaluation (mean 5.36, SD 2.14 vs mean 4.52, SD 2.14) and deduction (mean 9.88, SD 2.68 vs mean 7.27, SD 2.89). A mixed result		
Liu (2013)	A descriptive study	30 Chinese students majoring in sciences Random sampling Second-year undergraduate	The California Critical Thinking Skills Test (CCTST)-2000 designed by California Assessment Center (CAC) Translated to Chinese language	Chinese students' overall critical thinking skills scores (mean 19.83, SD 2.74) are higher than those of American students (mean 16.80, SD not specified). Chinese students demonstrate a lower level in inference (mean 7.37, SD 1.47) and induction skills (mean 7.18, SD 1.34), whereas other core skills including analysis (mean 4.93, SD 1.08), evaluation (mean 7.53, SD 1.72), deduction (mean 10.73, SD 1.91) are more proficient.	Not culturally neutral Small sample size	1*

				A mixed result		
Lun, Fischer, & Ward (2010)	A comparison, correlational study (Only consider the pilot study section because the main study includes a wider group: Asia students)	24 Chinese students and 35 New Zealand European students Not clarify the sampling strategy In university level (not specify undergraduate, postgraduate, or other levels)	Halpern Critical Thinking Assessment Using Everyday Situations (HCTAES) Only include the close-ended section of the HCTAES	Chinese students (mean -1.26, SD 1.70) perform worse than New Zealand European students (mean 0.87, SD 1.13) in the critical thinking test. Lower CT skills	Only focus on the skill dimension of critical thinking The paper-and-pencil form of assessment Only use one test to measure critical thinking	1*
Park, Niu, Cheng, & Allen (2021)	A correlational and cross-sectional study	166 Chinese and 103 American students The internet-based contact method (not specify the sampling strategy) In university level (not specify undergraduate, postgraduate, or other levels)	An updated Psychological Critical Thinking (PCT) Exam by Lawson et al. (2015) California Critical Thinking (CCT) Skills Test The experimental generation part from Sternberg Scientific Inquiry and Reasoning Averaged scores of these three tests: experiment generation (one vignette), PCT (two vignettes), and CCT (five sample items)	Chinese students (mean 1.32, SD 0.59) outperform American students (mean 1.02, SD 0.44) on critical thinking. Higher CT skills	Low level of representativeness of participants due to gender and discipline differences Only focus on three dimensions: evaluation, logical reasoning and probability thinking	1*

Zhang & Zhang (2013)	A correlational, cross-sectional study	197 Chinese students and 165 U.S. students The class-based contact method (not specify the sampling strategy) In university level (not specify undergraduate, postgraduate, or other levels)	Motivated strategies for learning questionnaire (MSLQ) from Pintrich et al (1991) Adopt the critical thinking subscale (the alpha reliability for U.S. 0.86) Translated to Chinese (the alpha reliability for Chinese 0.90)	Chinese students (mean 3.67, SD 0.92) perform better than U.S. students (mean 3.24, SD 0.87) in the critical thinking test. Higher CT skills	The instrument characteristics: developed in the U.S., likely to be inappropriate for Chinese students Self-report responses	1*
----------------------	--	--	---	--	---	-----------

Table 2. Studies on critical thinking dispositions (n=6)

Author(s) & date	Research design	Sample & level of education	Measuring instrument(s)	Finding(s) & result(s)	Limitation(s)	Rating
Dennett (2014)	A cross-sectional, comparative study	41 Chinese and 50 American students Voluntary sampling In both undergraduate and postgraduate levels	California Critical Thinking Disposition Inventory (CCTDI)	No significant differences in critical thinking dispositions are identified between Chinese and American students. No difference	Difficult to generalize because of the voluntary sampling Bias introduced by the researcher's experience of teaching Use only one instrument to measure critical thinking Closed-ended instrument, no space for alternatives Need to consider factors such as Chinese students' choice of studying abroad, prior experiences and university teachers' methods to develop critical thinking	1*
Lee et al. (2011)	A cross-sectional, comparative descriptive design	355 Korean students and 407 Chinese students in nursing education Stratified convenience sampling All levels of undergraduate	Critical thinking Scale developed by Yoon (2004) Translated to Korean language (Cronbach's alpha 0.85) Translated to Chinese language (Cronbach's alpha 0.81)	Chinese students demonstrate lower scores of critical thinking (mean 94.43, SD 7.26), compared to Korean students (mean 95.60, SD 8.59). Lower CT dispositions	Hard to control differences in nursing school systems, languages, and culture in these two countries Self-reported questionnaires	1*
McBride, Xiang, Wittenberg, & Shen (2002)	A cross-cultural, comparative, and descriptive study	218 American students and 234 Chinese students in physical education programmes	The California Critical Thinking Dispositions Inventory (CCTDI)	American students score higher in truth-seeking [mean 35.17 (from the table) /38.17 (from the text), SD 5.59 vs mean 34.62, SD	Hard to generalize because of the Chinese sampling strategy	1*

		<p>Selective sampling for American universities, and voluntary sampling for American students; purposive sampling for Chinese students</p> <p>Undergraduate: juniors or seniors</p>	<p>Translated to Chinese language (reliability coefficient 0.78)</p>	<p>5.65], inquisitiveness (mean 44.01, SD 8.91 vs mean 43.29, SD 5.80), maturity [mean 42.66, SD 6.75 vs mean 39.35 (from the table)/ 30.35 (from the text), SD 6.08] and self-confidence (mean 43.90, SD 6.69 vs mean 40.72, SD 6.02) than Chinese students.</p> <p>Lower CT dispositions</p>		
Petrini & Kawashima (2003)	A cross-sectional, comparative, descriptive study	<p>165 Japanese (82 students are 21-25 years old with no nursing related experiences; 83 students are with at least 5 years of experience), 300 Chinese (all are 21-25 years old and hardly have clinical experience) and 70 Samoa nursing students (all are 16-62 years old and with diverse nursing experience)</p> <p>Convenience sampling in each country</p> <p>Undergraduate</p>	<p>The California Critical Thinking Dispositions Inventory (CCTDI)</p> <p>Translated to Japanese (Cronbach's alpha 0.83) and Chinese languages (Cronbach's alpha 0.81)</p>	<p>A significant difference in critical thinking is evidenced between Japanese and Chinese students (Tukey's Honestly Significant Difference: $P < 0.05$). However, there is no difference between Chinese and Samoa students ($P > 0.05$, Tukey's Honestly Significant Difference: not specified).</p> <p>The total scores of CCTDI of Chinese students (mean 277.75, SD 23.18) are higher than Japanese (mean 271.84, SD 22.04).</p> <p>Chinese students show lower scores in truth-seeking (mean 31.38, SD 5.32 vs mean 34.87, SD 5.17), open-mindedness (mean 37.52, SD 4.73 vs mean 41.78, SD 4.15), inquisitiveness (mean 46.28, SD 5.77 vs mean 46.64, SD 5.48), and maturity (mean 36.93, SD 6.51 vs</p>	<p>Small sample size and convenience sampling, hard to generalise results</p> <p>Lack of some demographic information (e.g. educational background, admission criteria)</p> <p>The cultural-embedded instrument</p>	1*

				mean 43.73, SD 5.21), while higher in analyticity (mean 42.34, SD 5.38 vs mean 36.59, SD 4.48), systematicity (mean 38.84, SD 5.05 vs mean 35.13, SD 5.48) and self-confidence (mean 44.47, SD 6.04 vs mean 33.10, SD 7.51), compared with the Japanese cohort.		
				A mixed result		
Tiwari, Avery, & Lai (2003)	A cross-sectional, descriptive, and comparative study	222 Hong Kong Chinese students and 162 Australian nursing students Convenience sampling All levels throughout the pre-registration and post-registration nursing programme	The California Critical Thinking Inventory (CCTDI) Translated to Chinese language (Overall alpha 0.70)	Chinese students scored lower in all seven aspects: truth-seeking (mean 31.30, SD 4.52 vs mean 35.03, SD 6.94), open-mindedness (mean 38.40, SD 3.70 vs mean 41.86, SD 6.22), analyticity (mean 41.32, SD 4.12 vs mean 41.73, SD 6.01), systematicity (mean 37.13, SD 4.97 vs mean 38.51, SD 6.16), self-confidence (mean 40.27, SD 5.83 vs mean 40.74, SD 6.50), inquisitiveness (mean 43.60, SD 5.79 vs mean 46.29, SD 6.56), and maturity (mean 36.34, SD 5.29 vs mean 43.57, SD 6.74). Overall, Chinese students display a negative critical thinking disposition (mean 268.36, SD 21.58), whereas the Australian group are more inclined to positive ones (mean 287.73, SD 30.98).	Hard to generalize results because of the snapshot design, convenience sampling and high level of missing data	1*





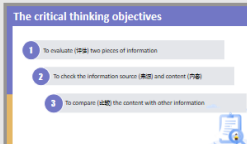
				Lower CT dispositions		
Yeh & Chen (2003)	A comparative, correlational, cross-sectional research design	214 nursing Chinese students in Taiwan and 196 nursing students in the USA Convenience sampling Undergraduate (juniors and seniors)	California Critical Thinking Dispositions Inventory (CCTDI) Translated to Chinese language (overall Cronbach's alphas 0.71)	Chinese students gain lower scores in six subscales including truth-seeking (mean 30.97, SD 4.86 vs mean 39.15, SD 6.29), open-mindedness (mean 40.90, SD 4.60 vs mean 43.90, SD 5.70), analyticity (mean 43.01, SD 4.09 vs mean 43.06, SD 5.50), systematicity (mean 38.28, SD 5.17 vs mean 41.11, SD 6.60), self-confidence (mean 42.47, SD 6.14 vs mean 42.94, SD 6.67) and maturity (mean 39.47, SD 5.14 vs mean 45.73, SD 6.96) except for the inquisitiveness (mean 48.42, SD 5.39 vs mean 47.34, SD 6.35). Overall, Chinese students show lower scores in critical thinking dispositions (mean 283.52, SD 21.39) than American undergraduates (mean 303.24, SD 29.38). Lower CT dispositions	Self-report critical thinking dispositions Convenience sampling Low level of generalisability due to the cross-sectional design Use different language versions of CCTDI	1*

Table 3. The study on critical thinking styles (n=1)

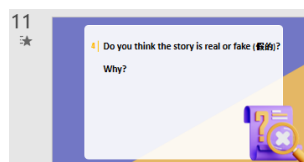
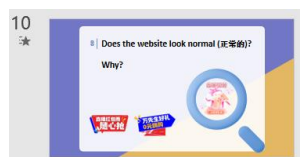
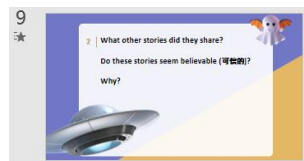
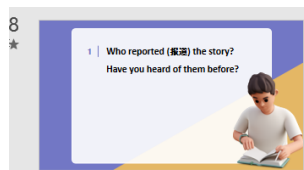
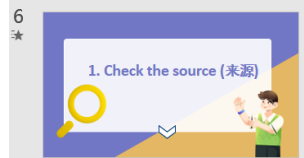
Author(s) & date	Research design	Sample & level of education	Measuring instrument(s)	Finding(s) & result(s)	Limitation(s)	Rating
Lu, Burris, Baker, Meyers, & Cummins (2021)	A cross-sectional study	104 U.S. students (37 males) and 103 Chinese students majoring in agriculture Convenience sampling Undergraduate	University of Florida Critical Thinking Inventory (UFCTI) Translated to a Chinese version (Overall reliability measured by the Cronbach's alpha 0.92)	Chinese students scored lower in engagement (mean 45.97, SD 10.19) than American students (mean 52.26, SD 6.25). Chinese students also scored lower in information seeking (mean 23.31, SD 5.30) than American students (mean 28.21, SD 3.55). U.S. students are more inclined to an engaging critical thinking style (mean 77.87, SD 5.05), whereas Chinese students prefer an information-seeking critical thinking style (mean 80.67, SD 4.96). [The overall scores are transposed and multiplied the engagement score by 1.866 due to the unequal number of items.] Information seeking	Using a convenience sample, limited in one university in each country, low level of generalizability Only exploring two constructs within critical thinking styles Only use one variable (country) to measure cultural differences	2*

Appendix C. An example of the teaching materials

Appendix C1. Lesson plan of lesson 5: Examine the information

<p>Lead-in (7 minutes)</p> <p>1 </p> <p>2 </p> <p>3 </p> <p>4 </p>	<p>Teacher: Let's look at a video from YouTube. YouTube is an online platform (网络平台) where anyone could post (发布) their videos, which is similar to Bilibili. The video is posted by a person called MrNuclearCat.</p> <p>After watching this video, you will answer my questions:</p> <ol style="list-style-type: none"> 1. What is this video about? 2. Do you trust (相信) this video? Why? <p>(Play the video on slide 3)</p> <p>(Students watching), and after 1 minute</p> <p>Teacher: What is this video about?</p> <p>(Students reply)</p> <p>Answer: An eagle (老鹰) snatches (一把抓起) a kid.</p> <p>Teacher: Do you trust it? Why?</p> <p>(Students reply)</p> <p>Note: If students reply "yes, I trust it." Teacher could ask the following prompt questions:</p> <ol style="list-style-type: none"> 1. Who is the author? Does the author use a real name? 2. Can anyone upload (上传) the video on the platform? 3. Do you notice the wing (翅膀) disappear at some point when the eagle flies in the sky? 4. Do you notice the strange shadow (奇怪的影子)? 5. Is an eagle able to seize (抓) a kid with that weight? <p>Teacher: In fact, this is a fake video (假视频) made by a student for his animation project (动画项目). The following video will explain how people find it fake.</p> <p>(Play the video on slide 4)</p>
<p>Critical thinking objectives (1 minute)</p> <p>5 </p>	<p>Teacher: In our daily life, there are many fake videos like this. In this lesson, you will learn how to evaluate information. Specifically, you need to evaluate the information source, content and compare it with other information.</p>

Presentation
(30 minutes)



1. Check the source

Teacher: Look at the website news.

Source: a website about UFOs, ghosts (鬼) and aliens (外星人)

Note: The website also published stories about a high-speed UFO captured on video (视频捕捉到高速不明飞行物) and ghosts caught on security camera (监控) in Japan.

Teacher: Let's check the source together. Who reported the story? Have you heard of them before?

(Students reply)

Answer: The author's name is: the honest author. The name seemed made up. Maybe most of us haven't heard of the website or the author before.

Teacher: What other stories did they share? Do these stories seem believable (可信的)? Why?

Answer: The website also has stories about UFOs and ghosts. They seem not believable because there is no scientific explanation and the video and pictures could be photoshopped (图像处理).

Teacher: Does the website look normal (正常的)? Why?

(Students reply)

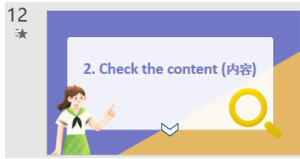
Answer: No. It has some commercial advertisements.

Teacher: After checking the information source, do you think the story is real or fake (假的)? Why?

(Students reply)

Answer: The story is more likely to be fake. (Students can give some reasons based on their answers to the above questions).

2. Check the content



Teacher: Look at the following short passage. We will check the content rather than fully trust it without evaluation. (Directly translate the passage if students do not understand the meaning)

An American study reported that people who have one egg a day are more likely to (更有可能) have heart disease (心脏病). This study included 30,000 adults with an average (平均) age of 52 in 2006. The researchers asked about their daily egg consumption (消耗), dietary habits (饮食习惯), and their personal information. Of the 30,000 people, 5,400 developed heart disease in 2023.

(1) Clarity: give clear information

Teacher: Does the research tell you how the eggs are prepared?

(Students reply)

Answer: No, the study did not mention the information.

Teacher: Why is it important to know how eggs are prepared?

(Students reply)

Answer: Eggs prepared in different ways may have a different influence on heart disease. For example, eggs could be boiled (水煮) and fried (煎).

(2) Accuracy: true facts

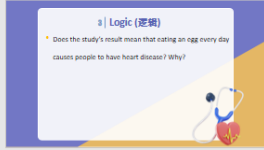
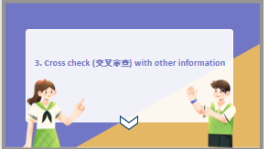
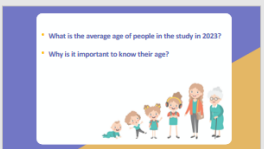
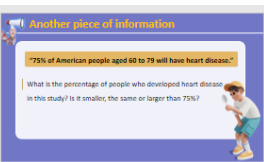
Teacher: How do the researchers know how many eggs people eat?


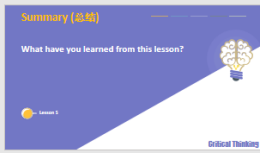
(Students reply)

Answer: They asked people. /People reported the information.

Teacher: Is this a reliable (可靠的) way of knowing the number of eggs people consume? Why?

(Students reply)

<p>16 ★</p> 	<p>Answer: No, it is not reliable because people self-reporting the situation may have a degree of subjectivity (主观性).</p> <p>Teacher: Can you suggest a more reliable way (更可靠的方式) of knowing the number of eggs people consume?</p> <p>(Students reply)</p> <p>Answer: People record their egg consumption through pictures or videos. / Researchers randomly (随机地) observe people's dietary (日常饮食). / ...</p> <p>(3) Logic</p> <p>Teacher: Does the study's result mean that eating an egg every day causes people to have heart disease? Why?</p> <p>(Students reply)</p>
<p>17 ★</p> 	<p>Answer: No. it only shows that eating an egg a day is correlated with heart disease. There may be other factors that cause heart disease such as age and lifestyle.</p> <p>3. Cross check (交叉审查) with other information</p> <p>Teacher: What is the average age of people in the study in 2023?</p>
<p>18 ★</p> 	<p>(Students reply)</p> <p>Answer: 69 years old. $[56+(2023-2006)=69]$</p> <p>Teacher: Why is it important to know their age?</p> <p>(Students reply)</p>
<p>19 ★</p> 	<p>Answer: At this age, there is already a high probability of heart disease. / Older people are more likely to have heart disease.</p> <p>Teacher: Now I give you another piece of information: "75% of American people aged 60 to 79 will have heart disease." What is the percentage of people who developed heart disease in this study? Is it smaller, the same or larger than 75%?</p>

<p>20 ★</p>  <p>An American study reported that people who have one egg a day are more likely to have heart disease. This study included 30,000 adults with an average age of 52 in 2006. The researchers asked about their daily egg consumption and dietary habits. Over the next 10 years, 4,400 developed heart disease.</p> <p>Does this study give strong evidence supporting that we should not eat eggs every day? Why?</p>	<p>(Students reply)</p> <p>Answer: The percentage is 18% ($5400/30000=18\%$). It is much smaller than 75%.</p> <p>Teacher: Does this study give strong evidence supporting that we should not eat eggs every day? Why?</p> <p>(Students reply)</p> <p>Answer: No. Because people in the study are on average 69 years old now. We do not know whether the result would be the same in other age groups. / Because the study does not ask people how the egg is prepared. / ...</p>
<p>Summary (2 minutes)</p> <p>21 ★</p>  <p>Summary</p> <p>What have you learned from this lesson?</p> <p>Critical Thinking</p>	<p>Teacher: What have you learned from this lesson? Please share your answers with desk mate.</p> <p>(Students talking), and after 1 minute,</p> <p>Teacher: Who'd like to share your ideas?</p> <p>(Students answer)</p>

Appendix C2. Student handout of Lesson 5: examine the information

Information 1. Check the source (来源): the website story



1. Who reported (报道) the story? Have you heard of them before?

2. What other stories did they share? Do these stories seem believable (可信的)? Why?

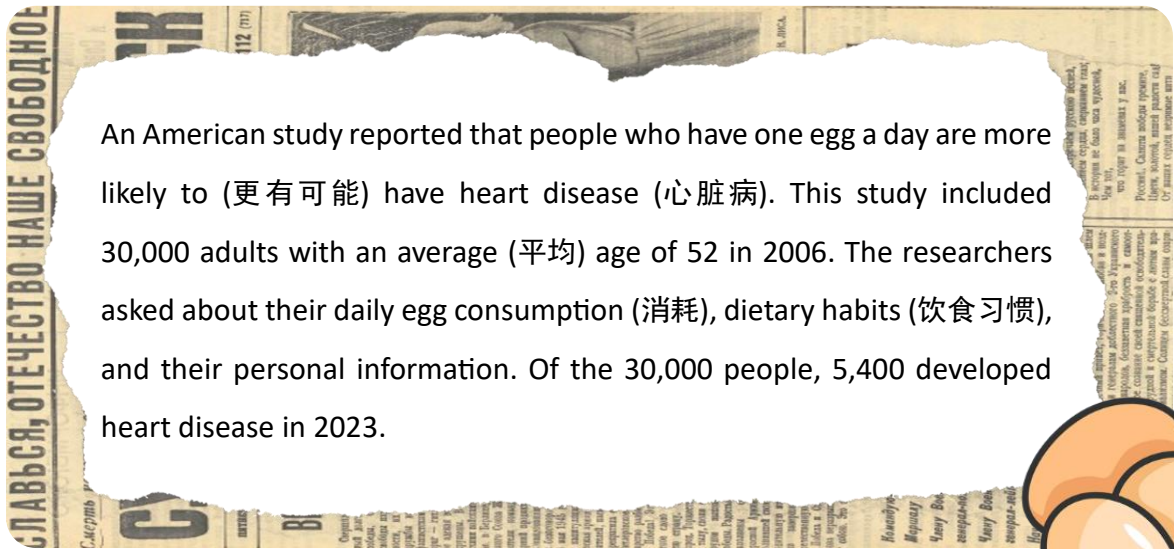
3. Does the website look normal (正常的)? Why?



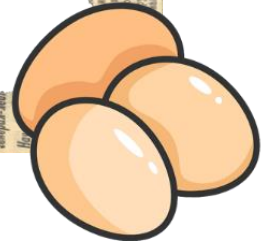
4. Do you think the story is real or fake (假的)? Why?



Information 2. Evaluate the content (内容) and cross check (交叉审查): the research report



An American study reported that people who have one egg a day are more likely to (更有可能) have heart disease (心脏病). This study included 30,000 adults with an average (平均) age of 52 in 2006. The researchers asked about their daily egg consumption (消耗), dietary habits (饮食习惯), and their personal information. Of the 30,000 people, 5,400 developed heart disease in 2023.



1. Clarity (清晰度): give clear information

1) Does the research tell you how the eggs are prepared?

2) Why is it important to know how eggs are prepared?

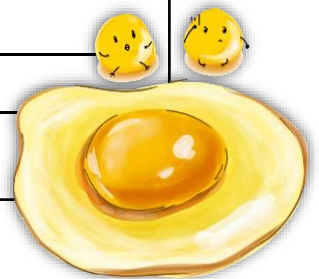


2. Accuracy (准确性): true facts

1) How do the researchers know how many eggs people eat?

2) Is this a reliable (可靠的) way of knowing the number of eggs people consume? Why?

3) Can you suggest a more reliable way (更可靠的方式) of knowing the number of eggs people consume?



3. Logic (逻辑)

1) Does the study's result mean that eating an egg every day causes people to have heart disease? Why?



4. Cross check (交叉审查) with other information

1) What is the average age of people in the study in 2023?

2) Why is it important to know their age?

3) Here is another piece of information: **“75% of American people aged 60 to 79 will have heart disease.”** What is the percentage of people who developed heart disease in this study? Is it smaller, the same or larger than 75%?



5. Does this study give strong evidence supporting that we should not eat eggs every day? Why?



Appendix D. The pre-and post-CT tests

Appendix D1. The pre-test of CT skill and student questionnaire

第一部分. 思维能力

第一节. 理由

在本节中，每个题目都有一个问题和一些理由。仔细考虑每一个理由，并判断哪个理由是符合逻辑，并且与题目直接相关的。你只能根据材料提供的信息选择答案，不要代入个人观点或常识。请把你的答案填在方框里，如你认为 A 是正确答案，请在方框里写 A。

1. 雇主是否应该允许所有员工选择灵活的工作时间？

- A. 是，重视员工的机构平均效益更高，而且员工流动率更低。
- B. 是，灵活的工作时间有利于员工平衡工作和生活，提高他们的工作效率。
- C. 否，如果所有员工都是兼职工作的话，那么公司的利润会降低。
- D. 否，员工应该按照合同里的规定进行工作。

2. 是否应该禁止互联网上的匿名发帖和匿名评论？

- A. 是，这可以减少网络暴力，因为使用真实身份的作恶者将被追究责任。
- B. 是，这可以减少网络暴力，因为人们将不再使用网络。
- C. 否，因为人们应该在互联网上自由发表评论。
- D. 否，因为在互联网上发布图片和评论不会伤害任何人。

3. 公司应该通过裁员来节约资金、实现利润最大化吗？

- A. 应该，在经济困难时期，裁员将使公司免于破产。
- B. 应该，公司没有义务雇用超出其能力范围的员工数量。
- C. 不应该，裁员会使员工士气低落，并导致生产力下降。
- D. 不应该，裁员会损害公司一直以来塑造的社会形象。

第二节. 假设

在本节中，每则材料之后都会有一些假设。默认每项假设本身是正确的，你需根据材料信息来判断哪项假设在材料中成立。请记住，你的答案只能基于材料信息。

4. 材料：周一到周五的上午九点至下午五点，妮娜通常会在办公室工作，但她已经安排了本周三下午三点去看牙医。

材料中的假设是：

- A. 妮娜周四不工作。
- B. 妮娜不用因为要去看牙医而请假。
- C. 妮娜已经请了本周三的假。

5. 材料：2008 年，美国总统承诺不会让国家陷入经济萧条的状况，但他失败了，因为在 2012 年初，超过 1200 万的美国公民失业了。

材料中的假设是：

- A. 2008 年的失业率高于 2012 年。
- B. 美国的失业率一直很高。
- C. 失业率是经济萧条的一项指标。

6. 材料：我看到两个戴着帽子的人在过马路，其中较矮的是一名女性。我之所以这么说，是因为当她摘下帽子时，我看到了她的长发。

材料中的假设是：

- A. 只有女性才有长发。
- B. 所有女性都很矮。
- C. 所有女性都有长发。

第三节. 推理

在本节中，每则材料之后都会有一些结论。你需要判断哪条结论能从材料中得出。你只能根据材料提供的信息选择答案，不要代入个人观点或常识。

7. 材料：自 2000 年以来，2012 年 5 月的降雨量已经达到了最高水平。有关降雨量的预报很少是准确的。

材料的结论是：

- A. 2012 年 5 月的降雨量超过了预期。
- B. 2012 年 5 月降雨量大于 2011 年 5 月。
- C. 五月通常是干燥的。

8. 材料：只有科技公司才能在 OTX (一个计算机安全平台)上市。没有哪一家科技公司能长期不稳定。

材料的结论是：

- A. 如果一家公司不是长期不稳定的，那么它会在 OTX 上市。
- B. 如果一家公司是在 OTX 上市的，那么它会长期不稳定。
- C. 如果一家公司是在 OTX 上市的，那么它不会长期不稳定。

9. 材料：统计数据显示，销售烘焙食品(如蛋糕和油酥糕点)的公司如果在广告中宣传自己来自法国或比利时，则更有可能获得成功。

材料的结论是：

- A. 以“法国”或“比利时”的名义宣传烘焙食品，更有可能带来更大的销量。
- B. 和其他国家的烘焙食品相比，法国和比利时的蛋糕和油酥糕点更贵。
- C. 和其他国家的烘焙食品相比，法国和比利时的蛋糕和油酥糕点质量更好。

第四节. 结论

在本节中，你将看到一则材料和相应的结论。这些结论不会直接出现在材料中，但它们可能从材料信息推断得出。每条结论之后，有五个选项供你选择：

- **完全正确**：如果你认为结论肯定是正确的。
- **可能正确**：如果你认为结论更有可能是正确而不是错误的，但没有足够的证据表明它肯定是正确的。
- **需要更多信息**：如果材料中的事实不能提供判断的依据。
- **可能错误**：如果你认为结论更有可能是错误而不是正确的，但没有足够的证据表明它肯定是错误的。
- **完全错误**：如果你认为结论肯定是错误的。

材料：两百名十几岁的学生主动参加了最近在英国伦敦举行的周末学生会议。在这次会议上，他们讨论了种族平等和如何实现世界和平的话题，因为这些都是学生们选出的当今世界最重要的问题。

10. 结论：这些学生来自英国各地。

根据题目中的材料，这一结论：

- A. 完全正确
- B. 可能正确
- C. 需要更多信息
- D. 可能错误
- E. 完全错误

11. 结论：和大多数十几岁的学生相比，参加这次会议的学生对广泛的社会问题表现出了更浓厚的兴趣。

根据题目中的材料，这一结论：

- A. 完全正确
- B. 可能正确

- C. 需要更多信息
- D. 可能错误
- E. 完全错误

12. 结论：在此之前，这些学生没有在他们的学校讨论过本次会议的话题。

根据题目中的材料，这一结论：

- A. 完全正确
- B. 可能正确
- C. 需要更多信息
- D. 可能错误
- E. 完全错误

第五节. 理解信息

在本节中，每则材料之后都会有一些结论。请默认这些材料信息本身是真实的，然后判断哪条结论符合材料信息。你只能根据材料提供的信息选择答案，不要代入个人观点或常识。

13. 材料：汉娜已经在伦敦市中心的一家律师事务所当了三年的事务律师。她希望能够升职。在汉娜的公司，员工要想升职，必须有至少四年的律师执业经验。

材料的结论是：

- A. 如果汉娜没有在三年内升职，那么对她现在的职位来说，她的资历太高了。
- B. 汉娜不能升职，因为她没有足够的律师执业经验。
- C. 我们不知道汉娜是否能升职。

14. 材料：每个被诊断出患有睡眠呼吸暂停症的人都曾与这个疾病作过斗争。例如，薇琪患了抑郁症，而且失去了工作；比尔感到他的婚姻关系有点紧张。

材料的结论是：

- A. 因为患有睡眠呼吸暂停症，薇琪失去了工作。
- B. 因为患有睡眠呼吸暂停症，比尔的婚姻失败了。
- C. 薇琪和比尔因为无法接受自己的疾病而进行了一场斗争。

15. 材料：最近，一本面向家长和老师的杂志发表了一份报告。该报告显示，吸烟的青少年在学校的成绩往往较低。随着吸烟次数的增加，学生的成绩在下降。该报告给出的一个建议是，我们可以通过禁止青少年吸烟来提高成绩。

材料的结论是：

- A. 我们支持该建议，因为研究发现吸烟会导致成绩下降。
- B. 我们支持该建议，因为研究发现减少吸烟会提高成绩。
- C. 我们不支持该建议，因为研究没有表明吸烟会导致成绩下降。

第二部分. 背景信息

这一部分会询问你的相关信息。你的答案将被匿名并保密，仅供本次研究使用，不会与你的老师和学校分享。

请根据个人情况在方框里填写 A 或 B。

16. 性别

- A. 男
- B. 女

17. 民族

- A. 汉族
- B. 少数民族

请在方框里填写数字。

18. 出生日期（年/月/日，如 2011/09/12）

/	/
---	---

19. 如果你家里有以下物件，请勾选 (√) “是”；如果无，请勾选 (√) “否”。

对于每一小题，请选择“是”或“否”，不要同时勾选。

	是	否
(a) 自己的房间		
(b) 用于学习的书桌		
(c) 可以用来做作业的电脑		
(d) 无线网络		
(e) 书架		
(f) 经典文学作品，如《红楼梦》		
(g) 诗歌集		
(h) 艺术品，如绘画作品		
(i) 关于艺术、音乐或设计的书		
(j) 乐器，如吉他，钢琴		

20. 在上一学年，父母与你进行以下活动的频率怎么样？请勾选 (√) 出最合适的选项。0 表示从未做过，10 表示经常做。

	0	1	2	3	4	5	6	7	8	9	10
(a) 讨论我在学校的表现											
(b) 辅导我的功课											
(c) 和我讨论政治或社会话题											
(d) 带我去图书馆或书店											
(e) 跟我讨论我在阅读的东西											

Part I. Thinking skills

Section 1. Arguments

In this section, each item has a question and a list of arguments for the answer. Consider which argument is logical, and directly related to the question. Base your answer only on the statement, not on your individual opinion or general knowledge. **Write the answer in the box. For example, if you think A is the answer, put A in the box.**

1. Should employers allow all staff the option of flexi-time working hours?

- A. Yes, organisations that value their staff are on average more productive and show lower staff turnover.
- B. Yes, giving employees flexi-time is good for their work-life balance, and their productivity.
- C. No, if all staff work part-time, the company will make little profit.
- D. No, workers should work according to what is in their contract.

2. Should anonymous posting and commenting on the internet be banned?

- A. Yes, this would reduce cyberbullying because perpetrators using their real identities would be held accountable.
- B. Yes, this would reduce cyberbullying because people would stop using the internet.
- C. No, because people should be free to comment on the internet.
- D. No, because posting images and comments on the internet does not harm anyone.

3. Should companies downsize their workforces to save money and maximise profits?

- A. Yes, downsizing will protect the company from bankruptcy in hard economic times.
- B. Yes, companies have no obligation to employ more people than they can afford.
- C. No, downsizing leads to the demoralisation of the workforce and causes reduced productivity.
- D. No, downsizing will damage the companies' social image, which they have been shaping for a long time.

Section 2. Assumptions

Each statement below is followed by a list of assumptions. Assume that each assumption is true. Based on the evidence in the statement, decide which of the assumptions follows the statement. Remember to base your response on the statement provided.

4. Statement: Nina usually works in the office from 9 am to 5 pm Monday to Friday, but she has scheduled an appointment with the dentist for 3 pm this Wednesday.

The assumption is that:

- A. Nina does not work on Thursday.
- B. Nina does not have to take days off for her dentist appointment.
- C. Nina has taken a day off on Wednesday.

5. Statement: In 2008, the president of the USA promised to prevent the country from entering an economic depression, but he failed because at the beginning of 2012, over 12 million USA citizens were unemployed.

The assumption is that:

- A. Unemployment is higher in 2008 than in 2012.
- B. Unemployment in the USA has always been very high.
- C. Unemployment is an indicator of economic depression.

6. Statement: I saw two people across the road wearing hats. The shorter of the two was a female. I say this because I saw her long hair when she removed her hat.

The assumption is that:

- A. Only females have long hair.
- B. All females are short.
- C. All females have long hair.

Section 3. Deductions

Each statement below is followed by a list of conclusions. Decide which conclusion follows the statement. You must select your answer based only on the information presented, and not on your opinion or general knowledge.

7. Statement: May 2012 had the highest level of rainfall on record since 2000. Predictions of rainfall are rarely accurate.

The conclusion is that:

- A. It rained more than expected in May 2012.
- B. The rainfall in May 2012 was greater than in May 2011.
- C. May is usually dry.

8. Statement: Only technological companies are listed on the OTX (a computer security platform). No technological company remains unstable for a long time.

The conclusion is that:

- A. If one company is not unstable for a long time, it will be listed on the OTX.
- B. If one company is listed on the OTX, it will be unstable for a long time.
- C. If one company is listed on the OTX, it will not be unstable for a long time.

9. Statement: Statistics have shown that companies selling baked goods (such as cakes and pastries) are more likely to be successful if they advertise themselves as being from France or Belgium.

The conclusion is that:

- A. Advertising baked goods as “French” or “Belgian” is more likely to result in more sales.
- B. Compared with baked goods from other countries, French and Belgian cakes and pastries are more expensive.
- C. Compared with baked goods from other countries, French and Belgian cakes and pastries are of better quality.

Section 4. Inferences

The statement below is followed by some inferences. These inferences do not appear directly in the statement, but can be inferred from the statement. There are five possible answers after each inference:

- **True**, if you believe the inference is definitely true.
- **Probably true**, if you think that it is more likely to be true than false, but there is not enough evidence to suggest that it is definitely true.
- **More information required**, if the facts from the statement provide no basis for the judgement.
- **Probably false**, if you think that it is more likely to be false than true, but there is not enough evidence to suggest that it is definitely false.
- **False**, if you believe the inference is definitely false.

Statement: Two hundred students in their early teens voluntarily attended a recent weekend student conference in London, England. At this conference, the topics of race equality and ways of achieving world peace were discussed, since these were the problems the students selected as being most important in today's world.

10. Inference: These students came from all parts of the UK.

Based on the statement, this inference is:

- A. True
- B. Probably true
- C. More information required
- D. Probably false
- E. False

11. Inference: The students who attended this conference showed a keener interest in broad social problems than most other students do in their early teens.

Based on the statement, this inference is:

- A. True

- B. Probably true
- C. More information required
- D. Probably false
- E. False

12. Inference: Prior to this, these students had not discussed the conference topics in their schools.

Based on the statement, this inference is:

- A. True
- B. Probably true
- C. More information required
- D. Probably false
- E. False

Section 5. Interpreting information

Each question below is a text followed by a series of conclusions. Assume that the information in the passage is true, and decide which conclusion follows the text. Based on your answer only on the text and not your opinion or general knowledge.

13. Text: Hannah has been working as a solicitor in a central London law firm for three years. She hopes to be promoted. To be promoted at Hannah's firm, employees must have at least four years' experience practising as a solicitor.

The conclusion is that:

- A. In three years' time, assuming that Hannah has not been promoted, she will be overqualified for her current position.
- B. Hannah cannot be promoted because she does not have sufficient experience of practising as a solicitor.
- C. We cannot know whether Hannah can be promoted or not.

14. Text: Everyone who has been diagnosed with sleep apnea has fought a personal battle owing to the disease. For example, Vicki suffered from depression and lost her job, while Bill felt a strain on his marriage.

The conclusion is that:

- A. Vicki lost her job because of her sleep apnea.
- B. Bill's marriage failed because of his sleep apnea.
- C. Vicki and Bill fought a personal battle because they could not come to terms with their disease.

15. Text: A recent report in a magazine for parents and teachers showed that adolescents who smoke cigarettes also tend to get low grades in school. As the number of cigarettes smoked increased, students' grades decreased. One suggestion made in this report was that we could improve students' grades by preventing adolescents from smoking.

The conclusion is that:

- A. The suggestion is supported because the research found that smoking causes grades to decrease.
- B. The suggestion is supported because the research found that reducing smoking can improve grades.
- C. The suggestion is not supported because the research does not show that smoking causes grades to fall.

Part II. Background information

This part asks for information about yourself. Your answers will be anonymised. Your responses here are for use in this research only and will be kept confidential. They will not be shared with your teacher or school.

Please write **A** or **B** in the box based on your personal situation.

16. Your birth sex

- A. Male

B. Female

17. Your ethnicity

A. Han

B. Minority

Please write down **numbers** in the boxes.

18. Date of birth (Year/Month/Day, e.g. 2011/ 09/12)

/	/
---	---

19. Please **tick (√)** **Yes** if the objects are in your home, and **tick (√)** **No** if the objects are not in your home. For each question, tick either Yes or No. Do not tick both.

	Yes	No
(a) A room of your own		
(b) A desk to study at		
(c) A computer you can use for homework		
(d) Wi-Fi		
(e) Bookshelves		
(f) Classic literature (e.g. <i>The Story of the Stone</i>)		
(g) Books of poetry		
(h) Works of art (e.g. paintings)		
(i) Books on art, music, or design		
(j) A musical instrument (e.g. a guitar or a piano)		

20. In the last academic year, how often did your parents do the following activities with you? **Tick (√)** the most appropriate response, from **0 (Never)** to **10 (All the time)**.

	0	1	2	3	4	5	6	7	8	9	10
(a) Discuss how I am doing at school											
(b) Help me with my homework											
(c) Discuss political or social issues											
(d) Take me to a library or bookstore											
(e) Talk to me about what I am reading											

Appendix D2. The post-test of CT skill

第一节. 理由

在本节中，每个测试题目都有一个问题和一些理由。仔细考虑每一个理由，并判断哪个理由是符合逻辑，并且与题目直接相关的。你只能根据材料提供的信息选择答案，不要代入个人观点或常识。请把你的答案填在方框里，如你认为A是正确答案，请在方框里写A。

1. 政府应该参与太空探索吗？

- A. 应该，因为我们需要考虑未来。
- B. 应该，因为太空探索的成果可以应用于工业，从而促进本国的经济发展。
- C. 不应该，政府已经在太空探索上花费了数万亿美元。
- D. 不应该，太空探索应该由私人公司进行。

2. 小学是否应该为孩子们提供学习编程的机会？

- A. 应该，孩子们在以后的工作和生活中将会用到编程技能。
- B. 应该，小学教育应该帮助孩子们找到兴趣点。
- C. 不应该，小学课程应该按照教育部门的教学大纲来设置。
- D. 不应该，把编程作为必修课会增加孩子们的学习压力。

3. 企业在培训员工方面的投资是值得的吗？

- A. 是，现在的员工们希望能够接受定期的培训。
- B. 是，研究表明，企业的培训支出越多，获得的利润就越多。
- C. 否，员工在工作中比在正规培训中学得更好。
- D. 否，员工在工作中往往会忽视他们在培训中学到的东西。

第二节. 假设

在本节中，每则材料之后都会有一些假设。默认每项假设本身是正确的，你需根据材料信息来判断哪项假设在材料中成立。请记住，你的答案只能基于材料信息。

4. 材料：为了节省时间，我们最好坐飞机去那里。

材料中的假设是：

- A. 坐飞机出行快。
- B. 坐飞机出行比坐火车方便。
- C. 坐飞机出行比搭乘其他交通工具快。

5. 材料：一开始，我们的项目计划从 5 月持续到 11 月，但因为暑假，我们申请延期到 12 月。

材料中的假设是：

- A. 我们的项目资金不足。
- B. 我们项目的工作人员放了暑假。
- C. 我们整个暑假都在一起。

6. 材料：这个村子里只有十支枪。我之所以这么说，是因为两位警戒人员手里各有一支枪，而且还有八支枪被堆放在村子中央。这就是我们所能看到的全部。

材料中的假设是：

- A. 村子里所有的枪支都在我们看得到的地方。
- B. 那八支堆放起来的枪已经上好了膛。
- C. 枪支是这些村民们唯一的武器。

第三节. 推理

在本节中，每则材料之后都会有一些结论。你需要判断哪条结论能从材料中得出。你只能根据材料提供的信息选择答案，不要代入个人观点或常识。

7. 材料：蒂莫西拥有一家成功的科技公司。在宣布重返办公室开展全职工作的计划后，他公司里 70%的员工要求每周至少有几天继续远程工作，而其中一小部分人决定离开公司。

材料的结论是：

- A. 所有科技公司的员工都喜欢远程工作。
- B. 蒂莫西公司的大多数员工仍然希望远程工作。
- C. 如果员工被要求重返办公室，他们将离开公司。

8. 材料：有些假期在下雨。所有的下雨天都很无聊。

材料的结论是：

- A. 如果不下雨，就不无聊。
- B. 有些假期很无聊。
- C. 所有的假期都不无聊。

9. 材料：柯利是一家只使用天然产品生产香薰蜡烛的公司。它反对在动物身上进行试验，也不会公司的任何产品中使用杀虫剂。

材料的结论是：

- A. 柯利公司的香薰蜡烛可能很贵。
- B. 柯利公司蜡烛的香味是从水果中提取出来的。
- C. 柯利公司的香薰蜡烛不太可能含有人造的固化剂。

第四节. 结论

在本节中，你将看到一则材料和相应的结论。这些结论不会直接出现在材料中，但它们可能从材料信息推断得出。每条结论之后，有五个选项供你选择：

- **完全正确**：如果你认为结论肯定是正确的。
- **可能正确**：如果你认为结论更有可能是正确而不是错误的，但没有足够的证据表明它肯定是正确的。
- **需要更多信息**：如果材料中的事实不能提供判断的依据。
- **可能错误**：如果你认为结论更有可能是错误而不是正确的，但没有足够的证据表明它肯定是错误的。
- **完全错误**：如果你认为结论肯定是错误的。

材料：美国国家公路交通安全管理局被誉为人民安全的守护者。它计划在校车上安装安全带，以减少安全隐患。提出这一想法的目的是确保当地和公立学校的孩子们在出行时的安全。而各地区的监管机构将负责安装安全带相关事宜。

10. 结论：美国国家公路交通安全管理局打算通过在校车上安装安全带来提高学校交通的安全性。

根据题目中的材料，这一结论：

- A. 完全正确
- B. 可能正确
- C. 需要更多信息
- D. 可能错误
- E. 完全错误

11. 结论：在校车上安装安全带一定能减少孩子们出行的安全隐患。

根据题目中的材料，这一结论：

- A. 完全正确
- B. 可能正确

- C. 需要更多信息
- D. 可能错误
- E. 完全错误

12. 结论：各地区的监管机构建议美国国家公路交通安全管理局探索更多的方法来提高校车的安全性。

根据题目中的材料，这一结论：

- A. 完全正确
- B. 可能正确
- C. 需要更多信息
- D. 可能错误
- E. 完全错误

第五节. 理解信息

在本节中，每则材料之后都会有一些结论。请默认这些材料信息本身是真实的，然后判断哪条结论符合材料信息。你只能根据材料提供的信息选择答案，不要代入个人观点或常识。

13. 材料：在英国，大英国家图书馆拥有最多的公共藏书。

材料的结论是：

- A. 英国某处可能拥有更多藏书。
- B. 英国某处可能拥有更多公共藏书。
- C. 大英国家图书馆不在英国境内。

14. 材料：一项关于儿童词汇量的研究表明，儿童的口语词汇量从八个月时的零个单词增加到六岁时的 2562 个单词。

材料的结论是：

- A. 儿童的口语词汇量在六岁以后继续增加。
- B. 儿童在学习走路期间，词汇量增长最慢。
- C. 在这项研究中，没有一个孩子在六个月大的时候学会说话。

15. 材料：我有一个九个月大的女儿。她通常乐意被放在床上，然后很快就睡着了。但每次她祖父母晚上来家里拜访时，我把她放到床上后，她就会哭，而且会持续一个小时。

材料的结论是：

- A. 如果我女儿的祖父母下午来家里拜访，她就会安静地入睡。
- B. 我女儿不肯睡觉是因为她祖父母晚上来家里拜访。
- C. 无法从材料得知为什么当我女儿的祖父母晚上来家里拜访时，她会不停地哭。

Section 1. Arguments

In this section, each test item has a question and a list of arguments for the answer. Consider which argument is logical, and directly related to the question. Base your answer only on the statement, not on your individual opinion or general knowledge. **Write the answer in the box. For example, if you think A is the answer, put A in the box.**

1. Should governments be engaging in space exploration research?

- A. Yes, because we need to think about our future.
- B. Yes, because findings of such research can be applied to industry, boosting the economy of the host country.
- C. No, because countries have collectively spent trillions of dollars on space exploration research already.
- D. No, space exploration should be taken by private companies.

2. Should primary schools offer young children the opportunity to learn to code?

- A. Yes, children will use coding skills in work and life when they get older.
- B. Yes, the primary education should help children find their interests.
- C. No, the primary school curriculum should follow the syllabus of the education department.
- D. No, making programming a required course would increase the pressure on children.

3. Is it worthwhile for a business to invest in training employees?

- A. Yes, employees expect to receive regular training these days.
- B. Yes, research shows the more money spent on training, the more profits will be gained.
- C. No, employees learn better on the job than in formal training.
- D. No, employees tend to ignore what they learn in training when doing their job.

Section 2. Assumptions

Each statement below is followed by a list of assumptions. Based on the evidence in the statement, decide which of the assumptions follow the statement. Remember to base your response on the statement provided.

4. Statement: We need to save time in getting there so we'd better go by plane.

The assumption is that:

- A. Travelling by plane is fast.
- B. Travelling by plane is more convenient than travelling by train.
- C. Travelling by plane is faster than other means of transportation.

5. Statement: In the beginning, our project was planned to run from May to November, but because of the summer holidays, we have asked for an extension until December.

The assumption is that:

- A. The project was underfinanced.
- B. People working on the project go on holiday in the summer.
- C. We spent the whole summer holiday together.

6. Statement: There are only 10 guns in this village. I know this because each of the two lookouts had one gun and eight guns were stacked in the middle of the village. That's all that could be seen.

The assumption is that:

- A. All the guns in the village are in plain sight.
- B. The eight stacked guns are loaded.
- C. Guns are these villagers' only weapons.

Section 3. Deductions

Each statement below is followed by a list of conclusions. Decide which conclusion follows the statement. You must select your answer based only on the information presented, and not on your general knowledge or opinion.

7. Statement: Timothy owns a successful tech company. After announcing the plan to return to their offices full time, 70% of his employees have requested to continue working remotely at least a few days a week, while a small percentage of them decided to leave the company.

The conclusion is that:

- A. Employees at all tech companies like working remotely.
- B. Most employees at Timothy's company still want to work remotely.
- C. If employees are required to return to their offices, they will leave the company.

8. Statement: Some holidays are rainy. All rainy days are boring.

The conclusion is that:

- A. If it is not raining, it is not boring.
- B. Some holidays are boring.
- C. All holidays are not boring.

9. Statement: Coley is a company that produces scented candles, using only natural products. Coley is against testing on animals and does not use pesticides in any of its products.

The conclusion is that:

- A. Coley's scented candles are likely to be expensive.
- B. The scent from Coley's candles is made from fruits.
- C. Coley's scented candles are unlikely to contain man-made setting agents.

Section 4. Inferences

Each statement below is followed by an inference, which is information that is not in the statement, but can be deduced/inferred from the statement. There are five possible answers:

- **True:** if you believe the inference is definitely true.
- **Probably true:** if you think that it is more likely to be true than false, but it is not definitely true.
- **More information required:** if the facts provide no basis for judging one way or the other.
- **Probably false,** if you think that it is more likely to be false than true, but there is not enough evidence to suggest that it is definitely false.
- **False,** if you believe the inference is definitely false.

Statement: The National Highway Traffic Safety Administration (NHTSA) of the United States of America is known as the protector of the safety of the people. The administration is aiming to promote the installation of seat-belts in school buses to reduce the safety hazards. The purpose of proposing this idea is to ensure that children in local and state schools are safe while traveling. Local regulators are the authority in carrying out the decision of executing the installation of seat-belts.

10. Inference: NHTSA intends to improve the safety of school transportation by proposing the idea of seat-belts.

Based on the statement, this inference is:

- A. True
- B. Probably true
- C. More information required
- D. Probably false
- E. False

11. Inference: The plan of seat-belts would definitely help in reducing the safety risks.

Based on the statement, this inference is:

- A. True
- B. Probably true
- C. More information required
- D. Probably false
- E. False

12. Inference: Local regulators have recommended NHTSA to explore more ways to enhance the safety of school buses.

Based on the statement, this inference is:

- A. True
- B. Probably true
- C. More information required
- D. Probably false
- E. False

Section 5. Interpreting information

Each question below is a text followed by a series of conclusion. Assume that the information in the passage is true, decide which conclusion follows the text. Based your answer only on the text and not your general knowledge.

13. The British National Library has the largest collection of publicly-owned books in the United Kingdom.

The conclusion is that:

- A. There might be a larger collection of books in the United Kingdom.
- B. There might be a larger collection of publicly-owned books in the United Kingdom.
- C. The British National Library is not in the United Kingdom.

14. A study of vocabulary growth in children from ages eight months to six years old shows that the size of spoken vocabulary increases from zero words at age eight months to 2,562 words at age six years.

The conclusion is that:

- A. Children's size of spoken vocabulary continues to increase after six years old.
- B. Vocabulary growth is slowest during the period when children are learning to walk.
- C. None of the children in this study had learned to talk by the age of six months.

15. I have a nine-month-old baby at home, and she usually agrees to be put to bed, where she falls asleep promptly. But every time her grandparents visit in the evening, she cries when I put her to bed, and she continues to cry for an hour.

The conclusion is that:

- A. My baby will go to sleep quietly if her grandparents visit in the afternoon.
- B. My baby refuses to go to sleep because of her grandparents' visit in the evening.
- C. We do not know why my baby continues to cry when her grandparents visit in the evening.

Appendix E. Teacher questionnaire

Infusing Critical Thinking in English Lessons

Thank you for participating in this important study on thinking skills conducted by Durham University, UK. Your response in this survey can help us understand the perceptions of different groups of teachers. It takes about 5 minutes to complete. You can use Chinese or English whichever suits you to complete.

This study has received ethical approval from the School of Education Ethics Committee of Durham University and is conducted in accordance with the British Educational Research Association's ethical guidelines. All answers from this survey are for use in this research only, will not be shared with any third party, and will be anonymised for reporting purposes. Your responses will be kept confidential. No school or individuals will be named or identified.

Completion of this survey is voluntary. You are free to withdraw at any time without giving a reason. By responding to this survey, you are agreeing to your anonymous responses and data being used as part of this project and to participate in the study.

We ask for your name to allow us to compare your answers now with the answers in a later survey. Your name will be removed once your responses are analysed and will not be identified in any reporting. If you have any questions regarding this survey or the project, please contact Keji Fan, email: keji.fan@durham.ac.uk

Your name

1. The following statements are about people's attitudes towards reading research findings in the newspaper or in a magazine. For each of the statement, indicate how much you agree with it. Please tick (✓) one response in each row from **0 (do not agree at all) to 10 (completely agree)**.

Statements:	0	1	2	3	4	5	6	7	8	9	10
The findings must be true if ...											
(a) the research is published recently.											
(b) the research is published in a reputable journal.											
(c) the research is conducted by a well-known scientist.											
(d) the research is supported by data collected using tests that are standardised and not developed by the researcher.											
(e) the research is conducted on a large number of people.											

2. For each statement below, indicate how much you agree with it. Please tick (✓) one response in each row from **0 (do not agree at all) to 10 (completely agree)**.

	0	1	2	3	4	5	6	7	8	9	10
(a) Critical thinking should be taught in school.											
(b) Critical thinking is not relevant to the English curriculum.											
(c) My teacher training courses prepared me well to teach critical thinking in the classroom.											
(d) Teachers need to have practical training to be able to teach critical thinking.											

3. In your English lessons, how often are students required to do the following activities?

Please tick (✓) one response in each row from **0 (never) to 10 (always)**.

	0	1	2	3	4	5	6	7	8	9	10
(a) Memorise facts and basic concepts.											
(b) Explain their answers.											
(c) Apply what they have learnt in new situations.											
(d) Think of alternative explanations (or reasons) for their answers.											
(e) Question the trustworthiness (可信度) of information received.											
(f) Create new ideas of their own.											

4. To what extent do you think the following are a barrier to teaching critical thinking in school? Please tick (✓) one response in each row from **0 (do not agree at all) to 10 (completely agree)**.

	0	1	2	3	4	5	6	7	8	9	10
(a) Emphasis on passing exams											
(b) Large class sizes											
(c) No clear definition of what critical thinking is											
(d) Students being not encouraged to question authority											
(e) Students' lack of background knowledge											
(f) Teachers' lack of training in introducing critical thinking in lessons											

If you think that there are other barriers, please indicate them below:

This section asks for your basic information, which may influence critical thinking skills of students. Please write your answers in the boxes.

5. Your age

6. Your birth sex:

- A. Male
- B. Female
- C. Prefer not to say

7. Educational background

7.1 Have you attended a normal university (师范院校)?

- A. Yes
- B. No

7.2 What is your highest educational qualification?

- A. Undergraduate
- B. Master
- C. Doctorate
- D. Other

If you choose D (other), please indicate your highest educational qualification:

7.3 Did you attend an overseas institution for your degree?

- A. Yes
- B. No

If you choose A (Yes), please indicate whether it is an English-speaking area or not:

8. Work experience

8.1 How many years have you been teaching English in secondary schools?

9. If there is anything else you would like to tell us, please write it down.

Appendix F. An example of class observation notes

Lesson 5 timetable (3rd April-7th April)

Date	School	Teacher	Class	Time
Monday (3 rd April)	D	16	3	9:10-9:55
Tuesday (4 th April)	C	17	1	9:20-10:00
Wednesday (5 th April)	A	18	8	14:10-14:50
		14	16	15:00-15:40
		6	14	15:50-16: 30
Thursday (6 th April)		3	22	8:20-9:00
7		19	15:00-15:40	
Friday (7 th April)		B	5	6
	2		8	9:30-10:10
	15		4	10:50-11:30
	A	1	21	15:50-16: 30

A summary of this lesson observation:

1. Teacher's instruction and translation could influence Ss' thinking. Some teachers did not translate the questions well, so students had no idea what to do.
2. For the research report about egg consumption and heart disease, most teachers instructed that the research was fake. However, it is real, just that we need to be cautious about applying the results to our daily life.
3. Most classes did not suggest better ways of knowing the number of egg people have. Only one or two students could think about using random sampling. I asked some students if they have learned sampling knowledge in their math class. They told me that it was last term's knowledge, but their math teacher said it would not be tested, so they did not learn it in class. This is common in China: If it is not tested, it would not be taught.

3 April 09:10-09:55am

School D Teacher 16 Class 03

location: the usual room

The observation includes: the class observation notes and recording, the informal talk with the teacher.

1. The lead in video (Awareness of evaluating the video)

Most Ss said that they trusted the video because: the toddler cries; the video recorded the whole process. But one student said that the video could be photoshopped. When the teacher asked some prompt questions like: what's the author's name? What is the platform? Could an eagle seize a toddler with that weight? Ss came to question the video content.

2. Lack discussion

The teacher did not ask Ss to discuss the website story and she guided through the questions quickly. Therefore, there are about 10 minutes left, so the teacher asked the whole class to review the content, and summarise what they've learned. More time should be needed for Ss to talk about what they've seen, and what information they can get from the website picture.

Also, some open questions should be discussed, but when one or two Ss gave their answers (most were expected from our lesson plan), the teacher explained and went to the next question. This works for quick-witted students, but most other Ss were still thinking when the teacher or their peers gave answers. Once they realised that they cannot follow, some lost interests and did other things. This could be evidenced by their answers to the final question: *Does this study give strong evidence supporting that we should not eat eggs every day? Why?* The active Ss who always follow teacher's guidance, could give excellent answer:

- *No, because there are different preparations for eggs, which has different nutrition value. The heart disease may not be caused by eggs. It could be caused by other nurture factors. (T: For example?) like their diets, daily routines, etc. They are old, and the older they are, the easier they may get heart disease. I forgot one, but I am sure I know another one. (T: why don't you sit down and*

think for a while?) (1 minute later) I remembered what I want to say. The way they get the information is questionable. It may be unreliable.

This answer indicated that the student really understood what we tried to teach them.

However, for some of others, their answers were not so good:

- *No, because eggs have nutrition.*
- *We should eat eggs in a moderate amount.*

3. Teacher could ask questions based on Ss' answer, help them to clarify their thinking

While the class interaction was mainly between teacher and quick-witted students, the teacher could ask further. For instance,

T: Can you suggest other ways to know the number of eggs people consume per day?

Ss: We can appoint some people to observe and record. We can watch the CCTV.

T: We have 3,0000 people, is it realistic to observe all?

Ss: We can randomly select some to observe.

T: How to do it?

Ss: Draw and do the stratified sampling.

When talking about that participants can upload the CCTV recordings to researchers, one student said, *“it may not be accurate, because it can be photoshopped.”*

After talking about the logical issue in the egg example, teacher asked students to think of other factors that could influence the heart disease. Ss could give many answers: *the genetic factor; the immunity; age; daily routine; ...*

4 April 09:20-10:00am

School C Teacher 17 Class 01

location: the usual room

The observation includes: the class observation notes and recording, the informal talk with the teacher.

1. The lead-in video

Ss were concentrated on the video. They liked watching videos. Most said that they trusted the video, but the teacher did not ask further. The teacher picked one student and asked some prompt questions such as the author of the video and the requirement of posting videos in the platform, Ss thought that the video had some problem.

2. Discuss the website picture

Two minutes were given to Ss to read and discuss the website information. Teacher asked Ss what they saw from the picture, most of their answers showed that they could get different information: *strange creature's appearance was captured; the creature was captured in Argentina; the appearance looks like a combination of camels, rabbits and horses; it walked in all fours; it was photographed in 2008; the creature was strange.*

After Ss going through the information, the teacher guided them to check the source.

Ss were able to explain their judgement that the story was fake:

- *S1: firstly, it has no scientific basis. Secondly, it was posted in informal website.*
- *S2: seeing is believing.*
- *S3: the picture was unclear.*

After different Ss sharing their ideas, the teacher summarised their answers again, making sure that all other Ss understand the content. However, it cost time, and sometimes seemed repetitive.

3. Not ask for explanation, instead, explains for Ss

When asked about why it is important to know how eggs were prepared, one student answered that because some types of preparation could cause heart disease, while some others did not. Teacher could have asked what types, letting Ss to give examples to make her arguments clearer. Likewise, she did **not ask why** they thought the way of asking people about egg consumption is unreliable.

When Ss gave their keywords of answers, the teacher **always explained for them**. Ss did not have many chances to clarify their thinking because the teacher was doing the

job for them. Besides, maybe sometimes the teacher's explanation is not what the Ss meant. After the class, I pointed out this issue to the teacher, and she agreed that maybe sometimes the students did not think as far as she. She knew this was a problem, but **it is not easy to change her teaching style within a short time. Besides, because of the teaching workload and testing pressure, teachers have to make good use of every minute, trying to explain for their Ss. Otherwise, they are worried that Ss do not understand the content.**

An example of explain for Ss was when one student said that the environment is also a factor influencing the heart disease, the teacher explained in class that if people live next to some factories that have much pollution, they are more likely to have heart disease. When she explained, one student in class said "*depression*". Maybe the student's meaning was that a depressive environment could make people feel mentally ill, and may cause the heart disease. While teacher's explanation delivered some knowledge to Ss, it was **not helpful** for Ss to improving their thinking (not only gave answers but also explains, gives supporting evidence), and sometimes was **away from** the genuine meaning of Ss.

4. Suggested ways of getting an accurate answer

- *Check the expense record*
- *Watch the CCTV*
- *Set up a website, and asking people to fill the forms*
- *Ask the participants' families or friends*

5. Directly gave answers when there was little time

After Ss gave their answers, the teacher explained a lot. Therefore, there was little time. When talking about the last question, is the research report strongly supported the idea that we should not have eggs every day? She directly gave answers, rushing to finish the content.

6. Teacher's view

- She said that she was more comfortable with the language use in the lessons. At the first few lessons, she thought that she had to use English all the time, but

now she knew when to use English and when to use Chinese. [She taught some new words and phrases when explaining the content; sometimes students could use what they learnt from usual English class to reply. e.g. “*seeing is believing*” is a new phrase that Ss recently learned. One used it when asked whether they believed the website story was fake.]

- Ss did not give answers like the pictures or the videos could be photoshopped in today’s lesson.

5 April 14:10-14:50

School A Teacher 18 Class 08

location: the usual room

The observation includes: the class observation notes and recording, the informal talk with the teacher.

1. the lead-in video

Ss judged the content based on their feelings or imaginations.

S1: This is absolutely impossible.

T: Why? What is your reason?

S1: emm... from my knowledge, animals cannot attach human.

S2: I feel this is true.

T: Why?

S2: the eagle was hungry, and it saw the family do not have guns. The baby was plump.

S3: I also feel this is true.

2. Ss did not have the awareness to clarify their answers or explain/give supporting evidence to their answers.

The teacher asked Ss to read the website picture and write down some answers on their handouts. Most told the story when asked what they know. While most could identify that this was not believable, they cannot give reasonable explanations. Then, the teacher asked them to discuss in groups. “*You need to explain your answers, give your reasons.*” teacher said.

Ss could find out problematic points such as:

- *The picture was unclear, the stories lacked scientific basis, we cannot photograph ghosts.*
- *This is an English website that has Chinese commercial advertisements.*
- *The publish date is 1 April, the April Fool's Day.*

However, more interesting discussion could be: **is all information on the April Fool's Day fake?**

3. Ss discussion

They had a heated discussion when the teacher asked them to consider questions. But when it comes to sharing their ideas, they were silent. Although teacher invited some Ss to show answers, some did not speak either. The teacher told me that this may be because Ss were afraid of giving wrong answers. In her another class, however, Ss were more active, and could give many interesting answers.

4. English language skill

The teacher let Ss read by themselves, and most Ss could understand the slide content. She encouraged Ss to use English in class. But in most cases, Ss answered and wrote down their ideas in Chinese.

5. Summary

In the summary section, teacher asked Ss to reflect what they've learned. Ss read through their handouts and wrote down their ideas. While some students' summarise were not really about this lesson (e.g. *I have learned that we need to find out the focus on information; our observation needs to be careful*), most could give a good summary, showing that they understood this lesson. E.g.

- *1. check the information source (author, area, website, etc.)*
- *2. check the content: clarity, accuracy, logic, and cross check*
- *3. summarise whether the information is real or fake*
- *(I learned how to) identify whether the information is real or fake. We need to read carefully, pay attention to the details, the content clarity and where is the information gained.*

- *We cannot believe in the information that is not verified, lacking in clarity and has wrong methods. Seeing is believing. We cannot blindly trust it. Even if the information is posted by the authority, we cannot fully trust it. We need to be brave to question*

5 April 15:00-15:40

School A Teacher 14 Class 16

location: the usual room

The observation includes: the class observation notes and recording, the informal talk with the teacher.

1. the lead in video

Trust: *because the video cannot be photoshopped; there is no signs of photoshop; there are lots of amazing things in the world.*

Not trust: *the video could be photoshopped. Based on life experience.*

The teacher did not ask prompt questions (author; platform; the video content), so Ss did not question the source aspect.

2. The website story: Ss could find out some problems, could give many points, but lack explanations

- *There is no official publication. If the story is true, the Argentina government should publish it, not this website.*
- *The advertisements are in Chinese.*
- *if it's true, the story should be reported many times.*
- *The story did not say the specific cities.*

Then, the teacher guided Ss to go through the questions about the source. Some Ss' answers were keywords or short sentences. The teacher did not ask them to explain or give an example.

3. Teach the website story quickly, there are some reasons.

1) Familiarity of content

If the teacher could be more familiar with the content, she would know the expected time of each section. Sometimes she messed up the three different questions.

2) Teachers' question asking skills

The teacher did not know how to ask Ss to think further, or give more reasonable explanations. She told me that Ss have already gave the expected answer, so she taught it quickly.

3) Lack discussion

There is no discussion in this part. While some active Ss could give good answers (mostly were those sit in front, the teacher did not walk through the class), others may be still thinking. if the teacher went through the content quickly and Ss cannot follow up, they would feel struggling.

4. Egg example: directly give answers; lack sharing; Ss cannot think of the good ways to obtain the egg consumption information

Some Ss thought that the stale eggs could lead to heart disease. They did not think from the way eggs prepared. Then, the teacher directly gave answers. While there are some class discussion, the teacher did not ask them to share ideas. When she heard the right answer, she moved on to the next. Ss did not suggest many good ways of knowing the egg number.

5. Teacher views

- She thought that Ss know the question intention/meaning, but sometimes they have difficulty explaining it clearly.
- She suggested that maybe I model one class, and then they can teach in their own classes.
- Our lessons were useful and interesting. Ss were active in class. But she did not used to the noisy class, and she cannot hear them clearly.

5 April 15:50-16:30

School A Teacher 6 Class 14

location: the usual room

The observation includes: the class observation notes and recording, the informal talk with the teacher. About 8 Ss were not listening.

1. the lead in video

Trust: *because the incident could happen in life.*

Not trust: *the video could be clipped. (T: How do you know? S: the slow motion); I have seen a similar video and the eagle cannot snatch a baby with that weight.*

The teacher did not point out the issue of author and the platform.

2. the website story: While Ss could give their points, teacher asked for reasons.

Sometimes the teacher asked further based on Ss' answer

Teacher stressed that Ss gave their answers should not based on their feelings, but on the evidence. They need to give convincing clues.

Ss first try to evaluate the website story:

- *I have a feeling that it's fake.*
- *The website is not reliable, and it has advertisements.*
- *The pictures were unclear.*

When talking about the author's name, Ss thought the name was not reliable.

T: what kind of name is more reliable?

Ss: the one verified with real names; the formal one.

In the end, Ss could give a conclusive answer, showing that the website story is fake.

Teacher summarised for Ss that they need to evaluate the information source from the author, website, and other aspects.

3. Egg: Ss could use previous learned knowledge from our lessons

1) After reading the research report, Ss could give good answers and reasons. E.g.

T: Should we have an egg every day? Why?

Most Ss: Yes.

T: Why? You can discuss for a while. Let me give you a hint. You can link to what we've learned before.

(Students discussion)

T: please give your reasons, not based on your feelings. OK, who'd like to say?

S1: The relationship between eggs and heart disease is only correlation. There are many other factors such as gene could influence the heart disease. Besides, the information is not clear.

T: Well done. She used what we have learned in previous lessons to explain. (Repeat the answer to make all class heard)

In addition this explanation, Ss also identified the age issue. The older people are, the more likely to have heart disease. One student suggested that the heart disease may be influenced by the Covid.

2) Suggested ways of knowing the egg consumption

- *Ask people to complete a form*
- *Choose the healthy people to attend the study*
- *Go to their homes to observe*
- *Post online questionnaire*

Some answers may be off track, this was because the teacher did not ask the question clearly, and did not mention the large number of people attending the study.

4. Summary

Ss said that they knew it's important to evaluate information, and summarised ways of evaluating information: source, content and cross check with other information. Teacher made it clear that Ss could use the previous four lesson's topic to evaluate the content.

6 April 08:20-09:00

School A Teacher 3 Class 22

location: the usual room

The observation includes: the class observation notes and recording, the informal talk with the teacher. 5 Ss took a sick leave.

1. the lead in video

Most Ss trust: the incident could happen because it's a beast; videos cannot be photoshopped.

The teacher did not ask the prompt questions that guide Ss to question the author and platform.

2. Group discussion before going through the questions

Ss were silent when the teacher asked them to discuss. The teacher thought this may be because Ss were sleepy in the morning. Also, some were shy and afraid of making mistakes. However, they can give some ideas when sitting in their seats, not raising hands.

Ss could find out some clues to judge the story is fake:

- *This is a paranormal phenomenon.*
- *This creature is not verified by people.*
- *The pictures were unclear; it may be photoshopped or projected.*

Teacher did not ask further when some answers required more explanation.

3. Ss could use previous learned knowledge from our lessons to evaluate the egg example

Before guiding Ss to judge the content, some could spot some issues:

- *The research did not clarify what kind of problems could be caused by eating eggs.*
- *The (average) age is 52. This is an old age.*
- *Maybe this is not caused by eggs, but because of their own illness, or other factors.*

4. Ss did **not give good ways** to know the egg consumption even though the teacher gave the hint that the study included 30,000 people. E.g.

- *Arrange these people to a place together, and have them eat eggs*
- *Set a CCTV*

- *Shoot video to record the egg consumption*

5. Some Ss could give a conclusive answer to the last question of the egg example. The class did not complete the summary section.

6. A good lesson is made by teacher and Ss working together.

Ss were not engaged, not react to the teachers' guidance or questions. They showed less interest to the video. All these made the teacher feel bad. She thought she did not teach this lesson well and gave some reasons: *Ss were sleepy or ill (more Ss took sick leave that afternoon); teacher did not explain clearly.*

Things were better in her another class. Ss were more active, and some answers really impressed the teacher. For instance, one student insisted that the video was fake as an eagle cannot seize a child with that weight. And there were some arguments in class. Also, she was surprised that one student immediately gave the average age of people which required a simple calculation.

6 April 15:00-15:40

School A Teacher 7 Class 19

location: the usual room

The observation includes: the class observation notes and recording, the informal talk with the teacher.

1. the lead in video

Some said they trust it, and some said not. Most gave reasons for not trusting it: *it was recorded (purposely); an eagle is unlikely to snatch a kid with that weight; it cannot be that coincident (Some argued it could be).*

Ss were concentrated when watching the videos.

2. Group discussion and sharing

Ss could look into the website information:

- *It was published in 2008.*

- *It talked about UFOs and aliens.*
- *The picture was unclear, and it looked like being photoshopped.*
- *The (strange) appearance of the creature*
- *There was a WeChat sign.*
- *The advertisement was about livestreaming which has not been developed at that time.*

Ss not only spotted the 1 April (The April Fool's Day), but also explained how the old date (2008) made the story unbelievable: *“The story was published for a long time. At that time, the technology was not well developed, and the information could not be spread in time, so there might be some misunderstandings.”*

Teacher could ask for more explanation. We reduce the content this week to give Ss and teacher more time to discuss. However, the teacher did not know how to guide Ss to have more discussion. She said that Ss have already mentioned some answers, and she had to slow down the teaching pace. In the usual English class, teachers always guide Ss to go through the content. They kept talking and explaining, and if the answer is given by Ss, they then show the answer and go to the next question/task. Maybe the teacher has got used this teaching style, and were at lost when more time was given to solve the fewer questions.

3. Ss' answers were on the track

After reading the egg example, teacher asked should we do not have eggs every day? Most Ss said no and gave their reasons: *the information was unclear; there could be other factors such as genes.* This indicated that Ss could link this to our previous lessons. Besides, Ss were active and they were comfortable to talk about their ideas in their seats.

4. Summary

Teacher linked the content to Ss daily life. And asked them: Do you trust the information immediately? Some yes and some no. Then, she guided Ss to evaluate the information: looking at the source, and publication date, check if the information is well-rounded, and if the content is true.

5. Teacher views

- The quality of Ss were quite different. It's difficult to pay attention to all of them.
- Sometimes when there is too much discussion, Ss may talk about other irrelevant issues.
- I discussed the class size with her, asking if she thought the big class size is a problem. She agreed with this point. She also stressed the variation of Ss level, suggesting that it may be better to have small class size whose Ss have similar level.

7 April 08:40-09:20

School B Teacher 5 Class 06

location: the usual room

The observation includes: the class observation notes and recording, the informal talk with the teacher and Ss.

1. the lead in video

Most trust and gave the reasons like *I feel it's true; the people's reaction in the video looks real.*

A few argued it was fake as eagles were afraid of people.

2. Review what have been learnt in previous lessons when introducing the CT objectives.

3. The website story

Ss read and do exercises by themselves, and then teacher translated and gave the answer. She did not ask further, simply asking Ss to write the answers, and use English to write simple sentences.

4. Egg example

(1) the importance of teacher's translation and guidance in class. [Teachers should be cautious about their instruction; not speak or act casually]

Teacher guided Ss to go through all questions but did not translate the question “How do the researchers know how many eggs people eat?” she gave wrong translation and guidance, which could mislead Ss’ thinking. Therefore, when it came to the next question, Ss did not know what to do.

(2) If it is not tested, it will not be taught

Ss’ suggestions for more accurate ways of knowing the egg consumption were not good. I wondered if Ss have learned the random sampling in their math classes, so I asked several Ss after class. They told me that the sampling knowledge should be taught last term. But their math teacher said that this part would not be tested, so it was not taught in class.

This is a common situation in China. Most teachers’ teaching is based on what will be tested. If it’s not tested, most of them will not teach in class.

5. Ss were active and could give their views without raising their hands

Ss were active when the teacher asked questions. However, the teacher only stood in the front, and talked to those who sat in the first few rows.

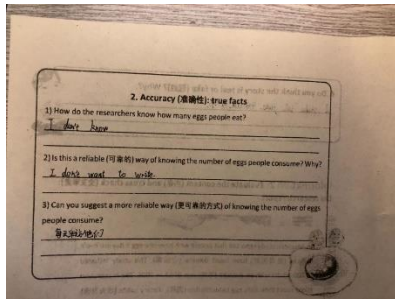
6. Write down answers in handout

After Ss gave their views, teacher gave the answer we suggested in lesson plan, and asked them to write them down. This was a common way of teaching in China when teachers explaining the test questions. As most test items have standard answers, teachers have to complete many teaching tasks, and the class time is limited, they give answers and Ss write down.

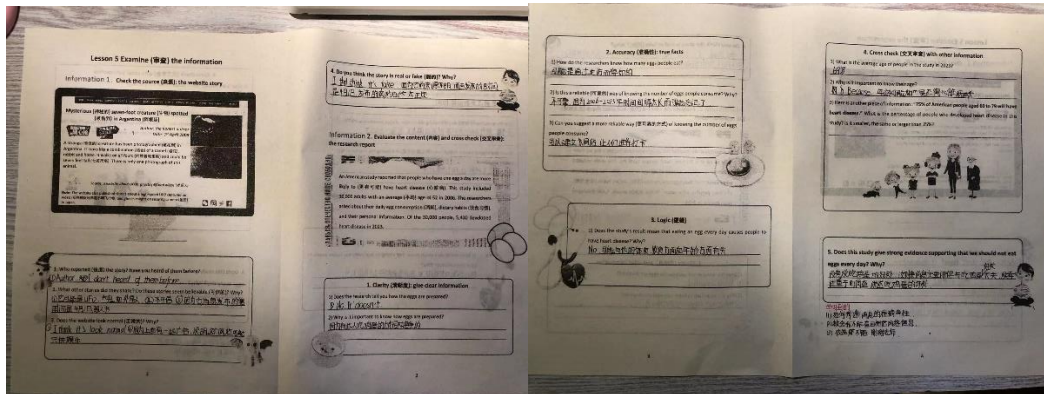
However, in our lessons, this way could prevent Ss from have their own thinking, as most were just waiting for the teacher’s answers. even if they gave the ideas, when their answers were different from their teachers’, Ss would mistakenly think that their answers were wrong, and teachers’ were correct.

Examples of Ss handouts

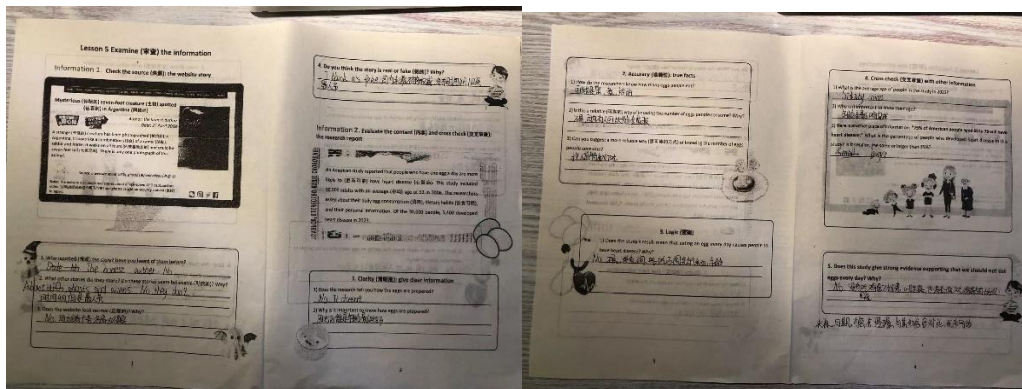
(1) Not listening or thinking



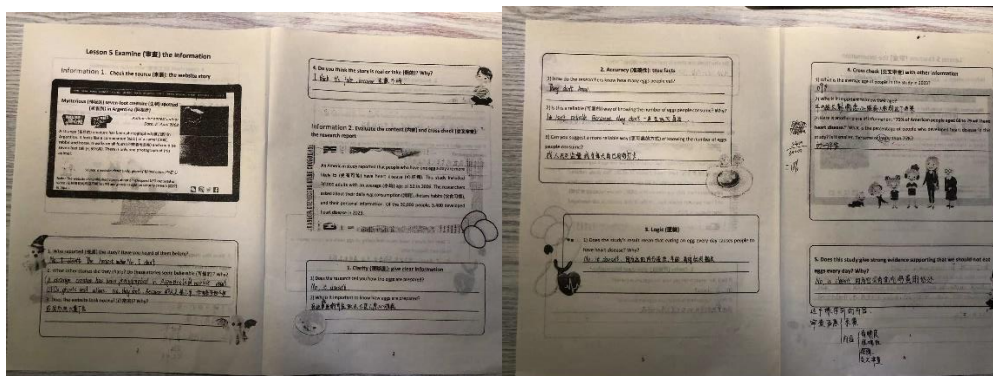
(2) The student thought he/she has learned a lot



(3) Ss use Chinese or English to complete the task

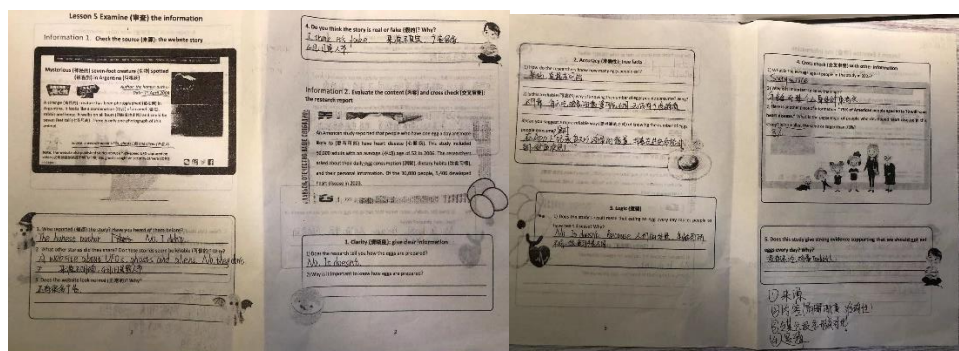


(4) Ss could draw a mind map to summarise the lesson content



(5) Ss did not share the answer about a more reliable way in class, but he/she was thinking and writing down the idea in handout.

Do a real-time record of the number of eggs consumed in App, and wrote down the health situation in physical examinations.



7 April 09:30-10:10

School B Teacher 2 Class 08

location: the usual room

The observation includes: the class observation notes and recording, the informal talk with the teacher.

1. the video

Before playing the video, the teacher helped Ss review the previous lessons. After watching the video, most Ss trusted the video, but the teacher did not ask them to give reasons.

2. the website story: Ss could spot some problems

The teacher kept asking “*Any other ideas?*” to ask Ss find out potential issues about the website picture. Below are some examples of Ss answers:

- *It has commercial advertisements; in general, formal website does not have advertisements.*
- *It's odd as these are Chinese advertisements, but this is an English website.*
- *The photo could be photoshopped.*
- *The WeChat below the story may induce people to click.*

3. Sometimes teacher directly gave answers

When one student raise hands and gave his answers, the teacher explained and directly talked about another point (there are different ways of egg preparation). She noticed

that earlier there was a student raising hands, so she asked the student to share his ideas. The student replied: “*That is what I want to say.*” (Students do not have the chance to share his ideas as the teacher has already talked about it.)

The conclusive questions in the website story and egg examples were designed to check if Ss could summarise the previous information examination, and give conclusive, well-round answers. However, the teacher may think these questions are repetitive, and directly gave the answers. Ss did not have the opportunity to evaluate the whole information, giving their judgement and explanations based the overall aspect.

4. Ss could give other ways to know the egg consumption, but their suggested ways were not good:

- *Do online questionnaires*
- *Use the lie detector (as they though the inaccuracy is due to people’s lie)*
- *Give rewards*
- *Assign people to observe*
- *Give some eggs to some people, and check the egg consumption after a period of time*
- *Set the CCTV*

5. Unclear instruction: too many pet phrases

The teacher used many pet phrases when she taught the lesson such as “*en... em... that’s to say ... this ... that ...*” This really influenced Ss’ learning. Also, she did not translate clearly. This may be due to her unfamiliarity with our lesson content though I saw she add many notes in her lesson plans. Also, this could be the teacher’s English skills. She cannot use English fluently and translate the words and sentences in the slides well.

7 April 10:50-11:30

School B Teacher 15 Class 04

location: the usual room

The observation includes: the class observation notes and recording, the informal talk with the teacher.

1. the video

Most trust the video because the recording is real. Then, the teacher asked the prompt questions to remind Ss of the author and the platform. Ss requested to see the video again. After watching the explanation video, Ss said “Oh!” together. Their reaction indicated that they understood the video was fake, and knew that they were fooled by the video too.

2. Discuss the website story

Before going through each question to check the source, teacher asked Ss to discuss “*What do you gain from the information I?*” Most Ss were engaged and had a heated discussion. But some were sleeping, not listening to the teacher. Their answers showed that they could notice different parts of the information.

- *A mysterious creature was photographed in Argentina; and the publication date was 2008’s April Fool Day.*
- *It looked like a combination of rabbits, camels, and horses.*
- *It was tall, and walked in all fours.*
- *It has advertisements.*
- *The picture*

When Ss shared their answers, Teacher reminded other Ss to take some notes. After going through the designed questions, Ss could give a conclusive answer to show that the story was fake.

3. Discuss the egg example

Ss were asked to read the research report in groups. Teacher asked “*Should we have eggs? Why?*” When Ss finished reading. Most said that we should have eggs, because:

- *Having eggs is likely to cause heart disease.*
- *People in the study were 52 years old.*

Before going through the questions to check the content, teacher guided Ss to review the previous four lessons, which were exactly the standards to examine the content.

Suggested ways of knowing the egg consumption:

- *Observe people every day*
- *Do a sampling survey*
- *Take pictures or videos*

In the end, Ss could give a good summary answer to show that the research report is not a strong evidence to support that we should not have eggs.

4. Clear instruction

Teacher could give a clear instruction and explanation in class. Therefore, Ss could summarise the lesson content and know how to apply it in their daily life. Teacher also linked this lesson to write research essays and do research.

7 April 15:50-16:30

School A Teacher 1 Class 21

location: the usual room

The observation includes: the class observation notes and recording, the informal talk with the teacher.

1. the lead in video

Ask Ss to discuss if they trust the video after watching it. Some trust: the baby was crying; the reaction of the people who recorded the video was real.

Some did not trust: it is only a video; this is too coincidental; it is like an incident in VR games.

2. Clear explanation

Teacher could give clear explanation in class, and also include some daily examples that Ss were familiar with, to raise Ss' awareness of the importance of examining the information.

3. Use the video example to help Ss review the previous lessons

4. Discuss the website story

Ss worked in groups and solved the questions. After discussion, Ss shared their answers, and teacher helped summarise the clues.

5. Discuss the egg example

Ss discussed, tried to solve the questions and shared in class. The whole class discussed actively. I listened to some groups' discussion, and found out that most of Ss' answers were on the track.

As for the question asking Ss to suggest better ways of knowing the egg number, they misinterpreted it to suggest better ways to do the whole research again (trying to find out the relationship between eggs and heart disease).

Appendix G. Ethical approval

Appendix G1. Ethical approval of the pilot study

Dear Keji,

The following project has received ethical approval:

Project Title: *Infusing critical thinking into Chinese secondary English curriculum: A pilot study*;

Start Date: *01 September 2022*;

End Date: *15 October 2022*;

Reference: *EDU-2022-04-01T23_18_11-qhtk28*

Date of ethical approval: *20 July 2022*.

Please be aware that if you make any significant changes to the design, duration or delivery of your project, you should contact your department ethics representative for advice, as further consideration and approval may then be required.

If you have any queries regarding this approval or need anything further, please contact ed.ethics@durham.ac.uk

If you have any queries relating to the ethical review process, please contact your supervisor (where applicable) or departmental ethics representative in the first instance. If you have any queries relating to the online system, please contact research.policy@durham.ac.uk.

Appendix G2. Ethical approval of the main trial

Dear Keji,

The following project has received ethical approval:

Project Title: *Infusing critical thinking into Chinese secondary English curriculum;*

Start Date: *01 January 2023;*

End Date: *01 July 2023;*

Reference: *EDU-2022-11-11T17_07_50-qhtk28*

Date of ethical approval: *14 November 2022.*

Please be aware that if you make any significant changes to the design, duration or delivery of your project, you should contact your department ethics representative for advice, as further consideration and approval may then be required.

If you have any queries regarding this approval or need anything further, please contact ed.ethics@durham.ac.uk

If you have any queries relating to the ethical review process, please contact your supervisor (where applicable) or departmental ethics representative in the first instance. If you have any queries relating to the online system, please contact research.policy@durham.ac.uk.

Appendix H. Sub-group analyses results

Appendix H1. Impact of EnglishFusion on critical thinking skills

Table 1 Sub-group analysis of CT skills by age

Younger students						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	497	7.68	8.95	1.27	2.56	0.19
Control	489	7.67	8.40	0.74	2.94	
Overall	986	7.67	8.68	1.01	2.77	
Older students						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	507	7.84	8.75	0.91	2.61	0.08
Control	518	7.62	8.31	0.69	2.86	
Overall	1025	7.73	8.53	0.80	2.74	

Table 2 Sub-group analysis of CT skills by birth sex

Girls						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	508	7.95	9.03	1.08	2.44	0.16
Control	500	7.85	8.50	0.66	2.86	
Overall	1008	7.90	8.77	0.87	2.66	
Boys						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	496	7.56	8.66	1.10	2.74	0.12
Control	507	7.44	8.21	0.77	2.94	
Overall	1003	7.50	8.43	0.93	2.85	

Table 3 Sub-group analysis of CT skills by ethnicity

Minority						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	17	8.00	8.41	0.41	2.98	0.32
Control	22	7.91	7.32	-0.59	3.17	
Overall	39	7.95	7.79	-0.15	3.09	

Majority/Han						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	987	7.76	8.86	1.10	2.58	0.13
Control	985	7.64	8.38	0.74	2.89	
Overall	1972	7.70	8.62	0.92	2.75	

Table 4 Sub-group analysis of CT skills by household objects

Own rooms						
Without a room						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	57	8.09	9.26	1.18	3.17	0.15
Control	62	7.85	8.56	0.71	2.99	
Overall	119	7.97	8.90	0.93	3.08	
With a room						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	947	7.74	8.82	1.08	2.55	0.14
Control	945	7.63	8.34	0.71	2.89	
Overall	1892	7.68	8.58	0.90	2.74	
Study desks						
Without a study desk						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	104	7.78	8.74	0.96	2.78	0.15
Control	103	7.55	8.11	0.55	2.71	
Overall	207	7.67	8.43	0.76	2.74	
With a study desk						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	900	7.76	8.86	1.10	2.57	0.13
Control	904	7.65	8.38	0.73	2.92	
Overall	1804	7.71	8.62	0.92	2.76	
Computers for homework						
Without a computer						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES

Experimental	765	7.74	8.84	1.10	2.56	0.14
Control	766	7.61	8.33	0.72	2.86	
Overall	1531	7.67	8.58	0.91	2.72	
With a computer						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	239	7.83	8.89	1.06	2.68	0.13
Control	241	7.76	8.45	0.69	3.03	
Overall	480	7.79	8.67	0.87	2.86	
Wi-Fi						
Without Wi-Fi						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	96	7.46	8.70	1.24	2.42	0.28
Control	112	7.54	8.01	0.47	2.90	
Overall	208	7.50	8.33	0.83	2.71	
With Wi-Fi						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	908	7.79	8.86	1.07	2.61	0.12
Control	895	7.66	8.40	0.74	2.90	
Overall	1803	7.72	8.63	0.91	2.76	
Bookshelves						
Without bookshelves						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	329	7.66	8.79	1.13	2.53	0.14
Control	322	7.59	8.34	0.75	2.88	
Overall	651	7.63	8.57	0.94	2.71	
With bookshelves						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	675	7.81	8.88	1.07	2.62	0.13
Control	685	7.67	8.36	0.70	2.91	
Overall	1360	7.74	8.62	0.88	2.78	
Classic literature						
Without classic literature						

	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	211	7.37	8.37	1.00	2.67	-0.08
Control	176	6.85	8.08	1.23	3.08	
Overall	387	7.13	8.24	1.11	2.86	
With classic literature						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	793	7.86	8.97	1.11	2.57	0.18
Control	831	7.81	8.42	0.61	2.85	
Overall	1624	7.84	8.69	0.85	2.73	
Books of poetry						
Without books of poetry						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	470	7.69	8.69	1.00	2.56	0.12
Control	402	7.46	8.14	0.68	2.91	
Overall	872	7.58	8.43	0.85	2.73	
With books of poetry						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	534	7.83	8.99	1.16	2.62	0.15
Control	605	7.76	8.50	0.74	2.90	
Overall	1139	7.79	8.73	0.94	2.78	
Works of art (e.g. paintings)						
Without works of art						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	650	7.68	8.83	1.15	2.61	0.14
Control	624	7.54	8.29	0.76	2.92	
Overall	1274	7.61	8.57	0.96	2.77	
With works of art						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	354	7.92	8.89	0.97	2.56	0.12
Control	383	7.81	8.46	0.64	2.87	
Overall	737	7.86	8.66	0.80	2.73	
Books on art, music or design						

Without books on art, music or design						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	698	7.77	8.87	1.11	2.52	0.16
Control	660	7.56	8.25	0.69	2.86	
Overall	1358	7.66	8.57	0.91	2.70	
With books on art, music or design						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	306	7.75	8.79	1.05	2.75	0.10
Control	347	7.80	8.56	0.76	2.98	
Overall	653	7.78	8.67	0.89	2.88	
Musical instruments (e.g. pianos or guitars)						
Without an instrument						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	752	7.71	8.85	1.14	2.61	0.18
Control	720	7.55	8.20	0.65	2.92	
Overall	1472	7.63	8.53	0.90	2.78	
With an instrument						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	252	7.90	8.85	0.95	2.53	0.03
Control	287	7.88	8.76	0.88	2.84	
Overall	539	7.89	8.80	0.91	2.70	

Table 5 Sub-group analysis of CT skills by degree of parental involvement

Discuss school performance with children						
Lower degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	550	7.61	8.76	1.15	2.49	0.18
Control	545	7.58	8.24	0.66	2.90	
Overall	1095	7.60	8.50	0.91	2.71	
Higher degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	454	7.94	8.95	1.02	2.71	0.09

Control	462	7.72	8.49	0.77	2.90	
Overall	916	7.83	8.72	0.89	2.81	
Help children with homework						
Lower degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	613	7.71	8.83	1.11	2.60	0.13
Control	622	7.61	8.35	0.74	2.94	
Overall	1235	7.66	8.59	0.92	2.78	
Higher degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	391	7.83	8.88	1.05	2.58	0.14
Control	385	7.70	8.37	0.67	2.83	
Overall	776	7.76	8.63	0.86	2.71	
Discuss political or social issues						
Lower degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	520	7.66	8.73	1.07	2.60	0.13
Control	552	7.60	8.32	0.72	2.92	
Overall	1072	7.62	8.52	0.89	2.78	
Higher degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	484	7.87	8.98	1.11	2.59	0.15
Control	455	7.70	8.40	0.70	2.87	
Overall	939	7.79	8.70	0.91	2.74	
Go to a library or bookstore together						
Lower degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	593	7.65	8.76	1.11	2.67	0.16
Control	556	7.68	8.34	0.66	2.95	
Overall	1149	7.66	8.55	0.89	2.82	
Higher degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES

Experimental	411	7.92	8.98	1.06	2.48	0.10
Control	451	7.60	8.38	0.78	2.83	
Overall	862	7.75	8.67	0.91	2.67	
Discuss children's reading						
Lower degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	553	7.64	8.75	1.11	2.58	0.14
Control	591	7.64	8.37	0.73	2.91	
Overall	1144	7.64	8.55	0.91	2.76	
Higher degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	451	7.90	8.97	1.07	2.61	0.14
Control	416	7.65	8.34	0.69	2.88	
Overall	867	7.78	8.67	0.88	2.75	

Table 6 Sub-group analysis of CT skills by overall parental involvement

Lower degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	496	7.57	8.73	1.17	2.60	0.16
Control	530	7.59	8.31	0.72	3.03	
Overall	1026	7.58	8.51	0.94	2.84	
Higher degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	508	7.95	8.96	1.01	2.58	0.12
Control	477	7.70	8.41	0.70	2.76	
Overall	985	7.83	8.69	0.86	2.67	

Table 7 Sub-group analysis of CT skills by schools

School A						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	604	7.62	8.70	1.09	2.59	0.18
Control	528	7.66	8.24	0.58	2.92	

Overall	1132	7.64	8.49	0.85	2.76	
School B						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	315	8.09	9.19	1.10	2.57	0.06
Control	303	7.77	8.70	0.93	2.99	
Overall	618	7.93	8.95	1.02	2.78	
School C						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	49	8.10	9.27	1.16	2.57	0.19
Control	104	7.64	8.32	0.67	2.59	
Overall	153	7.79	8.62	0.83	2.58	
School D						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	36	6.83	7.69	0.86	2.95	0.01
Control	72	6.96	7.78	0.82	2.76	
Overall	108	6.92	7.75	0.83	2.81	

Table 8 Intervention impact on overall CT skills of different academic achievers

Chinese subject						
Lower achievers						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	547	7.48	8.31	0.83	2.60	0.02
Control	326	6.92	7.69	0.77	2.83	
Overall	873	7.27	8.08	0.81	2.69	
Higher achievers						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	453	8.11	9.50	1.39	2.56	0.25
Control	674	8.00	8.69	0.69	2.94	
Overall	1127	8.04	9.01	0.97	2.81	
Maths subject						
Lower achievers						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES

Experimental	510	7.46	8.26	0.80	2.62	0.13
Control	347	7.22	7.66	0.44	2.83	
Overall	857	7.36	8.02	0.65	2.71	
Higher achievers						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	489	8.08	9.46	1.38	2.52	0.18
Control	652	7.88	8.75	0.87	2.94	
Overall	1141	7.97	9.06	1.09	2.77	
English subject						
Lower achievers						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	576	7.51	8.35	0.85	2.66	0.11
Control	370	7.20	7.76	0.56	2.85	
Overall	946	7.39	8.12	0.74	2.74	
Higher achievers						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	424	8.12	9.52	1.40	2.46	0.22
Control	632	7.91	8.71	0.80	2.93	
Overall	1056	7.99	9.03	1.04	2.77	

Table 9 Sub-group analysis by prior critical thinking levels

Lower critical thinkers						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	438	6.05	8.52	2.47	2.29	0.08
Control	469	5.70	7.98	2.28	2.56	
Overall	907	5.87	8.24	2.37	2.44	
Higher critical thinkers						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	566	9.08	9.10	0.02	2.29	0.28
Control	538	9.33	8.68	-0.65	2.45	
Overall	1104	9.21	8.90	-0.31	2.39	

Appendix H2. Impact of EnglishFusion on academic attainment

Table 10 Sub-group analysis of academic attainment by age

Younger students						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	491	200.49	198.55	-1.94	21.28	-0.06
Control	483	236.80	236.14	-0.66	21.67	
Overall	974	218.50	217.19	-1.31	21.47	
Older students						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	496	202.41	199.64	-2.77	22.62	-0.13
Control	511	234.77	234.89	0.12	20.86	
Overall	1007	218.83	217.53	-1.30	21.78	

Table 11 Sub-group analysis of academic attainment by birth sex

Girls						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	503	210.33	208.02	-2.30	20.37	-0.15
Control	493	239.66	240.44	0.78	20.27	
Overall	996	224.84	224.07	-0.78	20.37	
Boys						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	484	192.24	189.82	-2.41	23.51	-0.05
Control	501	231.92	230.64	-1.28	22.14	
Overall	985	212.42	210.58	-1.84	22.82	

Table 12 Sub-group analysis of academic attainment by ethnicity

Minority						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	17	186.71	184.94	-1.76	18.01	-0.08
Control	22	237.82	237.57	-0.25	21.81	
Overall	39	215.54	214.63	-0.91	20.00	
Majority/Han						

	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	970	201.71	199.35	-2.37	22.03	-0.10
Control	972	235.71	235.45	-0.26	21.25	
Overall	1942	218.73	217.42	-1.31	21.66	

Table 13 Sub-group analysis of academic attainment by household items

Own rooms						
Without a room						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	57	212.45	209.96	-2.48	24.62	0.06
Control	62	243.52	239.65	-3.87	19.61	
Overall	119	228.64	225.43	-3.21	22.07	
With a room						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	930	200.78	198.43	-2.35	21.80	-0.11
Control	932	235.24	235.22	-0.02	21.34	
Overall	1862	218.03	216.85	-1.18	21.60	
Study desks						
Without a study desk						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	99	195.06	196.52	1.46	25.76	0.02
Control	100	208.79	209.68	0.89	23.41	
Overall	199	201.96	203.14	1.17	24.55	
With a study desk						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	888	202.17	199.39	-2.78	21.47	-0.11
Control	894	238.77	238.39	-0.39	21.00	
Overall	1782	220.53	218.95	-1.58	21.26	
Computers for homework						
Without a computer						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	751	200.93	198.92	-2.01	21.95	-0.12

Control	758	231.74	232.24	0.50	21.27	
Overall	1509	216.41	215.66	-0.75	21.64	
With a computer						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	236	203.11	199.65	-3.46	21.99	-0.04
Control	236	248.65	245.95	-2.70	21.06	
Overall	472	225.88	222.80	-3.08	21.51	
Wi-Fi						
Without Wi-Fi						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	91	191.19	188.23	-2.96	23.18	0.04
Control	111	234.16	230.31	-3.85	19.56	
Overall	202	214.80	211.35	-3.45	21.22	
With Wi-Fi						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	896	202.50	200.20	-2.30	21.84	-0.11
Control	883	235.96	236.15	0.19	21.42	
Overall	1779	219.11	218.05	-1.06	21.66	
Bookshelves						
Without bookshelves						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	321	195.63	194.36	-1.27	22.47	-0.07
Control	316	219.36	219.74	0.37	23.08	
Overall	637	207.40	206.95	-0.45	22.77	
With bookshelves						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	666	204.26	201.38	-2.88	21.70	-0.11
Control	678	243.40	242.85	-0.55	20.35	
Overall	1344	224.00	222.30	-1.71	21.06	
Classic literature						
Without classic literature						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES

Experimental	207	176.88	173.25	-3.63	23.43	-0.12
Control	172	205.89	205.10	-0.78	22.36	
Overall	379	190.04	187.70	-2.34	22.96	
With classic literature						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	780	207.98	205.96	-2.02	21.55	-0.09
Control	822	242.01	241.86	-0.15	21.02	
Overall	1602	225.44	224.38	-1.06	21.30	
Books of poetry						
Without books of poetry						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	461	193.23	190.71	-2.52	23.06	-0.15
Control	394	220.31	221.26	0.95	21.89	
Overall	855	205.71	204.79	-0.92	22.58	
With books of poetry						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	526	208.66	206.45	-2.22	20.97	-0.06
Control	600	245.90	244.85	-1.05	20.80	
Overall	1126	228.51	226.91	-1.60	20.88	
Works of art (e.g. paintings)						
Without works of art						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	638	196.88	194.82	-2.06	23.19	-0.12
Control	616	228.90	229.43	0.53	21.00	
Overall	1254	212.61	211.82	-0.79	22.17	
With works of art						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	349	209.82	206.92	-2.90	19.53	-0.07
Control	378	246.94	245.39	-1.55	21.62	
Overall	727	229.12	226.92	-2.20	20.64	
Books on art, music or design						
Without books on art, music or design						

	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	686	200.30	198.60	-1.70	21.88	-0.10
Control	649	229.14	229.61	0.47	21.87	
Overall	1335	214.32	213.67	-0.65	21.89	
With books on art, music or design						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	301	204.08	200.23	-3.85	22.09	-0.11
Control	345	248.21	246.58	-1.63	20.00	
Overall	646	277.65	224.99	-2.66	21.01	
Musical instruments (e.g. pianos or guitars)						
Without an instrument						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	740	200.62	198.94	-1.68	22.14	-0.09
Control	708	230.86	231.13	0.27	21.08	
Overall	1448	215.41	214.68	-0.73	21.64	
With an instrument						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	247	203.95	199.57	-4.38	21.33	-0.13
Control	286	247.87	246.32	-1.55	21.64	
Overall	533	227.52	224.66	-2.86	21.52	

Table 14 Sub-group analysis of academic attainment by different degree of parental involvement

Discuss school performance with children						
Lower degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	537	197.17	190.63	-2.54	22.05	-0.15
Control	536	228.11	228.84	0.73	21.76	
Overall	1073	210.62	209.72	-0.91	21.96	
Higher degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	450	211.34	209.21	-2.13	21.87	-0.03

Control	458	244.71	243.29	-1.41	20.60	
Overall	908	228.17	226.40	-1.77	21.23	
Help children with homework						
Lower degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	602	200.50	198.37	-2.13	21.77	-0.08
Control	617	232.68	232.30	-0.38	21.83	
Overall	1219	216.79	215.55	-1.24	21.81	
Higher degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	385	202.95	200.23	-2.72	22.28	-0.12
Control	377	240.79	240.73	-0.06	20.29	
Overall	762	221.67	220.27	-1.40	21.34	
Discuss political or social issues						
Lower degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	512	196.80	196.05	-0.75	22.69	-0.02
Control	545	231.91	231.69	-0.22	22.08	
Overall	1057	214.90	214.43	-0.47	22.37	
Higher degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	475	206.47	202.38	-4.09	21.03	-0.18
Control	449	240.43	240.12	-0.31	20.22	
Overall	924	222.97	220.72	-2.25	20.72	
Go to a library or bookstore together						
Lower degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	583	196.47	194.26	-2.21	23.24	-0.09
Control	549	226.95	226.84	-0.10	22.02	
Overall	1132	211.25	210.06	-1.19	22.67	
Higher degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES

Experimental	404	208.65	206.07	-2.57	19.99	-0.11
Control	445	246.63	246.18	-0.45	20.28	
Overall	849	228.55	227.09	-1.46	20.16	
Discuss children's reading						
Lower degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	540	196.56	194.67	-1.88	22.35	-0.07
Control	582	227.59	227.21	-0.37	21.95	
Overall	1122	212.65	211.55	-1.10	22.14	
Higher degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	447	207.37	204.44	-2.93	21.49	-0.14
Control	412	247.30	247.20	-0.09	20.25	
Overall	859	226.52	224.95	-1.57	20.94	

Table 15 Sub-group analysis of academic attainment by overall parental involvement

Lower degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	485	191.90	190.14	-1.76	22.65	-0.09
Control	527	227.16	227.48	0.32	22.70	
Overall	1012	210.26	209.59	-0.68	22.69	
Higher degree of parental involvement						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	502	210.69	207.75	-2.94	21.27	-0.10
Control	467	245.46	244.55	-0.91	19.49	
Overall	969	227.44	225.48	-1.96	20.45	

Table 16 Sub-group analysis of academic attainment by schools

School A						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	590	183.26	180.71	-2.55	21.11	-0.24
Control	523	244.09	246.34	2.25	19.22	

Overall	1113	211.84	211.55	-0.30	20.38	
School B						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	314	241.32	235.94	-5.38	19.90	0.14
Control	301	252.20	244.00	-8.20	19.77	
Overall	615	246.65	239.88	-6.76	19.87	
School C						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	48	193.48	206.99	13.51	26.80	0.13
Control	99	182.68	192.81	10.13	25.97	
Overall	147	186.20	197.44	11.23	26.20	
School D						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	35	161.49	167.81	6.33	33.23	0.22
Control	71	178.65	179.12	0.47	23.68	
Overall	106	172.98	175.39	2.41	27.18	

Table 17 Intervention impact on different levels of students' academic attainment

Lower academic achievers						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	540	150.51	150.93	0.41	23.81	-0.12
Control	337	160.37	163.80	3.43	25.95	
Overall	877	154.30	155.87	1.57	24.68	
Higher academic achievers						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	447	262.99	257.29	-5.70	18.98	-0.19
Control	657	274.43	272.28	-2.15	18.11	
Overall	1104	269.80	266.21	-3.59	18.54	

Table 18 Intervention impact on higher and lower critical thinkers' academic attainment

Lower critical thinkers						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES

Experimental	427	185.26	183.23	-2.03	21.90	-0.08
Control	460	224.23	223.92	-0.32	21.83	
Overall	887	205.47	204.33	-1.14	21.87	
Higher critical thinkers						
	N	Pre mean	Post mean	Gain mean	Gain SD	Gain ES
Experimental	560	213.80	211.19	-2.61	22.02	-0.11
Control	534	245.69	245.48	-0.21	20.76	
Overall	1094	229.36	227.93	-1.44	21.44	