

Linguistic relativity and L2 acquisition: classifier sensitivity in intermediate Mandarin learners

Laura Friedrich and Tanja Kupisch

Underlying syntax of long-distance wh-questions in L2 English: insights from prosody

Tokitada Hangai and Shigenori Wakabayashi

The role of explicitness of L1 tense and aspect markers in L2 tense acquisition: examining the cline of difficulty hypothesis

Yuka Fujii

From eye-tracking to input design: enhancing verb processing in L3 Spanish for Chinese learners

Hui-Chuan Lu, An Chung Cheng, and Zi-Xu Liu

Students' experiences with generative AI in source-based writing: perceptions, concerns and mediational strategies

Taimin Tammy Wu and Chen-Yu Chester Hsieh

Guided use of Gen-AI for research article reading: a case study of EFL students in a Taiwanese MA-TEFL program

Chia-Hui Chiu

Input and perfuming: an interdisciplinary dialogue between Krashen's theory of second language acquisition and the Yogācāra cognitive model

Wei-Wei Zhao

Tailoring speaking activities: the impact of differentiated instruction on first-year university students

Adriani Yulia Purwaningrum

Integrating collaborative action research, problem-based learning, and team-based learning in a cross-national CLIL context: a case study in international marketing management

Chih-Cheng Lin

Developing non-native EFL teachers' empathy through learning a new foreign language

Linh Thuy Nguyen

Lecturers' beliefs about national language policies in English medium instruction programs in Vietnam

Tran Tan Phat

Taiwan Journal of Linguistics Book Series

ISBN 978-626-7886-19-9



9 786267 886199

Current Issues in L2 Teaching and Research Vol. 2

Current Issues in L2 Teaching and Research Vol. 2



Edited by

Stano Kong, Miao-Chin Chiu, Yu-Hsin Huang,
Kai-Lin Wu, Pin-Hua Wu, Yu-Han Lin and One-Soon Her

Current Issues in L2 Teaching and Research, Vol. 2

Edited by
Stano Kong
Miao-Chin Chiu
Yu-Hsin Huang
Kai-Lin Wu
Pin-Hua Wu
Yu-Han Lin
One-Soon Her

CONTENTS

PREFACE	i
LIST OF EDITORS	iv
LIST OF CONTRIBUTORS	vi

LINGUISTIC RELATIVITY AND L2 ACQUISITION: CLASSIFIER SENSITIVITY IN INTERMEDIATE MANDARIN LEARNERS	1
---	----------

Laura Friedrich and Tanja Kupisch

UNDERLYING SYNTAX OF LONG-DISTANCE WH- QUESTIONS IN L2 ENGLISH: INSIGHTS FROM PROSODY	40
--	-----------

Tokitada Hangai and Shigenori Wakabayashi

THE ROLE OF EXPLICITNESS OF L1 TENSE AND ASPECT MARKERS IN L2 TENSE ACQUISITION: EXAMINING THE CLINE OF DIFFICULTY HYPOTHESIS	88
--	-----------

Yuka Fujii

FROM EYE-TRACKING TO INPUT DESIGN: ENHANCING VERB PROCESSING IN L3 SPANISH FOR CHINESE LEARNERS	113
--	------------

Hui-Chuan Lu, An Chung Cheng, and Zi-Xu Liu

STUDENTS' EXPERIENCES WITH GENERATIVE AI IN SOURCE-BASED WRITING: PERCEPTIONS, CONCERNS AND MEDIATIONAL STRATEGIES	148
---	------------

Taimin Tammy Wu and Chen-Yu Chester Hsieh

GUIDED USE OF GEN-AI FOR RESEARCH ARTICLE READING: A CASE STUDY OF EFL STUDENTS IN A TAIWANESE MA-TEFL PROGRAM	186
---	------------

Chia-Hui Chiu

CONTENTS

- INPUT AND PERFUMING: AN INTERDISCIPLINARY
DIALOGUE BETWEEN KRASHEN'S THEORY OF SECOND
LANGUAGE ACQUISITION AND THE *YOGĀCĀRA*
COGNITIVE MODEL** 218

Wei-Wei Zhao

- TAILORING SPEAKING ACTIVITIES: THE IMPACT OF
DIFFERENTIATED INSTRUCTION ON FIRST-YEAR
UNIVERSITY STUDENTS** 254

Adriani Yulia Purwaningrum

- INTEGRATING COLLABORATIVE ACTION RESEARCH,
PROBLEM-BASED LEARNING, AND TEAM-BASED
LEARNING IN A CROSS-NATIONAL CLIL CONTEXT: A
CASE STUDY IN INTERNATIONAL MARKETING
MANAGEMENT** 284

Chih-Cheng Lin

- DEVELOPING NON-NATIVE EFL TEACHERS' EMPATHY
THROUGH LEARNING A NEW FOREIGN LANGUAGE** 323

Linh Thuy Nguyen

- LECTURERS' BELIEFS ABOUT NATIONAL LANGUAGE
POLICIES IN ENGLISH MEDIUM INSTRUCTION
PROGRAMS IN VIETNAM** 359

Tran Tan Phat

PREFACE

Welcome to *Current Issues in L2 Teaching and Research Vol. 2*, a collection arising from the second Tunghai International Conference on Second Language Teaching and Research, held on December 6–7, 2025. Following the success of the inaugural conference in 2023, the 2025 meeting was likewise jointly organized by the Department of Foreign Languages and Literature, the International Graduate Program of Teaching Chinese as a Second Language, and the English Language Center of Tunghai University.

The conference is held every two years, and an edited volume is published the following year. The papers collected here were drawn from presentations at the 2025 conference. As with the inaugural volume, the submissions were selected through double-blind review in accordance with the protocols of the *Taiwan Journal of Linguistics*; the program committee members served as editors.

In its broad sense, a second language (L2) is any language a person uses or studies in addition to the first language (L1), and L2 use has now become commonplace worldwide. The chapters that follow gather the pedagogies and research findings shared at the 2025 conference, with the aim of deepening our understanding of L2 learning and teaching and of prompting further discussion and collaboration among scholars. The volume is organized around three themes: acquisition and processing, generative AI in the L2 classroom, and theory and pedagogy.

Chapters 1 through 4 address acquisition and processing. In Chapter 1, Laura Friedrich and Tanja Kupisch ask whether Mandarin classifiers shape the conceptualization of German learners of L2 Mandarin. Comparing similarity ratings across L1 Mandarin, L2 Mandarin, and L1 German speakers, they find no overall classifier or gender effect; within the L2 group, however, an interaction with proficiency and exposure suggests a heightened—and possibly temporary—sensitivity to classifiers at the intermediate stage. In Chapter 2, Tokitada Hangai and Shigenori Wakabayashi use an Elicited Production Task to examine long-distance wh-questions produced by Japanese learners of English. Because Intonational Phrase boundaries reflect clause-level organization, prosody reveals that target-like word order can conceal more than one underlying derivation—variation that surface form alone would not detect.

In Chapter 3, Yuka Fujii administers a Temporal Interpretation Choice Task to Japanese and Chinese learners of English. The two groups reach comparable accuracy but rely on different strategies: Japanese learners depend primarily on morphological cues, whereas Chinese learners draw more heavily on context. Chapter 4, by Hui-Chuan Lu, An Chung Cheng, and Zi-Xu Liu, synthesizes three eye-tracking studies of how Chinese-speaking learners of L3 Spanish process verb morphology while reading. Learners lean on lexical and contextual cues at the expense of verb forms, and longer fixation on verb regions is associated with more accurate comprehension—though the authors caution that eye movements index attention rather than awareness or acquisition.

Chapters 5 and 6 turn to generative AI. In Chapter 5, Taimin Tammy Wu and Chen-Yu Chester Hsieh examine how 14 Taiwanese EFL students draw on GenAI during source-based writing. Interviews and textual analysis of drafts show that students use ChatGPT and ChatPDF mainly to understand source texts, verify information, organize ideas, and support citation decisions, while remaining wary of inaccurate output; they also rely on peer discussion, teacher feedback, instructional materials, library databases, and the writing center. AI-assisted writing, the authors argue, is best understood as a mediated activity in which learners coordinate technological, social, and institutional resources. In Chapter 6, Chia-Hui Chiu combines Move Analysis instruction with two GenAI tools, NotebookLM and ChatGPT, in an MA-level course on reading L2 English research articles. Students rated the Move Analysis instruction as helpful. Although the difference between the two tools was not statistically significant, the patterns suggest that NotebookLM offers more structured support for comprehension and synthesis, whereas ChatGPT lends itself to clarification and on-demand questions.

The remaining chapters take up theory and pedagogy. In Chapter 7, Wei-Wei Zhao reads Stephen Krashen's Monitor Theory alongside the cognitive model of Buddhist Yogācāra philosophy, proposing that acquisition may be understood in relation to the storehouse consciousness (*ālaya-vijñāna*), and learning and monitoring in relation to functions of the mental consciousness (*mano-vijñāna*). Yogācāra, the chapter argues, offers a complementary philosophical vocabulary for describing language internalization. Chapters 8 and 9 are pedagogical. In Chapter 8, Adriani Yulia Purwaningrum reports classroom action research with 40

Indonesian first-year students of L2 English, showing that speaking tasks differentiated by proficiency level and learning preference improve fluency, accuracy, and confidence while reducing speaking anxiety. In Chapter 9, Chih-Cheng Lin integrates Problem-Based Learning, Team-Based Learning, and Collaborative Action Research in an English-medium international marketing course taught in collaboration with a partner institution in the United States. Taiwanese and American students alike reported gains across all four dimensions of Coyle's 4C framework, yet only content understanding and communication were significantly associated with semester grades—a gap between conventional assessment and broader competence.

In Chapter 10, Linh Thuy Nguyen follows six EFL teachers in Vietnam as they learn a new foreign language. Through focus groups, diaries, semi-structured interviews, and thematic analysis, the author finds that the experience reactivates the teachers' empathy for students and prompts them to revise their teaching. In Chapter 11, Tran Tan Phat examines the beliefs of 10 Vietnamese lecturers about English medium instruction (EMI). The lecturers view EMI positively but temper that view against students' English proficiency, institutional support, and disciplinary demands; faced with increased workload, the competing claims of content and language, and an absence of clear guidelines, they use Vietnamese strategically to support comprehension.

We hope this volume lays a foundation for further inquiry among L2 teachers and researchers, and we look forward to the third conference in 2027 and the volume that will follow in 2028.

Stano Kong
Miao-Chin Chiu
Yu-Hsin Huang
Kai-Lin Wu
Pin-Hua Wu
Yu-Han Lin
One-Soon Her
July 24, 2026

EDITORS

Stano Kong is Lifetime Distinguished Professor in the Department of Foreign Languages and Literature and Founding Director of the International Graduate Program of Teaching Chinese as a Second Language (TCSL) at Tunghai University, Taiwan. He serves as Editor-in-Chief of the Journal of Chinese Language Teaching. His research focuses on generative approaches to third language (L3) acquisition, particularly the acquisition of French, Spanish, German, and Japanese by L1 Chinese speakers with advanced L2 English. He has published extensively in international journals on multilingual acquisition and currently investigates cross-linguistic transfer and syntactic development in L3 grammars.

Miao-Chin Chiu is an Associate Professor in the English Language Center at Tunghai University, Taiwan. She has over 25 years of experience teaching Chinese and English as second languages. Drawing on her extensive classroom experience, her research focuses on L2 teacher education, presentation skills, speaking anxiety, and the application of cognitive linguistics to second language pedagogy.

Yu-Hsin Huang is an Associate Professor in the International Graduate Program of Teaching Chinese as a Second Language at Tunghai University. Her research interests include second language acquisition, Chinese syntax, and generative linguistics. She has published widely in international journals and presented her work at major conferences. Her current research examines the acquisition of Chinese and English syntax from a cross-linguistic perspective, supported by several competitive research grants.

Kai-Lin Wu is an Assistant Professor in the Department of Foreign Languages and Literature at Tunghai University, Taiwan. Her research interests include second language writing, formative assessment, and AI-assisted writing, with a particular focus on their pedagogical applications. She teaches academic writing in English at the undergraduate and graduate levels and is the author of two English

composition textbooks for EFL learners in Taiwan.

Pin-Hua Wu is an Assistant Professor in the International Graduate Program of Teaching Chinese as a Second Language at Tunghai University, Taiwan. Her research interests include Chinese language teaching, instructional materials and pedagogy, Chinese lexical semantics, pragmatics, and second language acquisition. Her recent research examines learner language, Chinese lexical semantics, and their applications to Chinese language teaching.

Yu-Han Lin is an Assistant Professor at the English Language Center, Tunghai University. She has taught English and Chinese in Taiwan, Hong Kong, the United States, and online since 2007. Her research interests include the use of L2 Chinese, Taiwanese, and English for interaction in classrooms, eldercare settings, and television programs.

One-Soon Her is Nan Lin & Alice Lin Chair Professor of the Department of Foreign Languages and Literature at Tunghai University and an adjunct Chair Professor at the Graduate Institute of Linguistics, National Chengchi University. He is the founder and editor of *Taiwan Journal of Linguistics*, a Diamond Open Access journal. His recent research interests include numeral classifiers, numeral systems, typology of questions, and language policies. Most of his publications can be downloaded from his homepage at <https://onesoonher.github.io/info/index.html>.

CONTRIBUTORS

An Chung Cheng is Professor of Spanish at the University of Toledo. Her research focuses on second and third language acquisition, eye-tracking, psycholinguistics, language pedagogy, and technology-enhanced language learning.

Chia-Hui Chiu is currently an Associate Professor in the Department of Foreign Languages and Literature at Tunghai University. Her current research interests include the integration of artificial intelligence (AI) into EFL teaching and learning.

Laura Friedrich is a PhD student in the Department of Linguistics at the University of Tübingen, Germany. Her research interests include first and second language acquisition, psycholinguistics, and pragmatics. Her current work focuses on children's acquisition of verbal irony, with a particular focus on the role of common ground in irony comprehension.

Yuka Fujii is a master's student in Arts at University of Tsukuba, Japan, specializing in Second Language Acquisition. She holds a BA in Arts from the same institution. Her research investigates the acquisition of English tense and aspect by East-Asian learners, with a particular focus on Japanese speakers. Drawing on a semantic perspective, she explores the underlying difficulties learners face in acquiring the tense-aspect domain.

Tokitada Hangai received his B.A. and M.A. from Chuo University under Professor Wakabayashi's supervision. He is currently a Ph.D. student, and his research interests lie in Generative Approaches to Second Language Acquisition, particularly wh-questions and relative clauses.

Chen-Yu Chester Hsieh is currently an Assistant Professor in the Department of Foreign Languages and Literature at National Tsing Hua University. His research interests include pragmatics, discourse

analysis, corpus linguistics, and applied linguistics. He has published on a wide range of topics, including pragmatic expressions in Mandarin conversation, metaphor awareness and competence in second language learning, discourse competence in speaking assessments, and evaluative language in English-teaching YouTubers' videos.

Tanja Kupisch is Professor of Linguistics at Lund University, Sweden. Her research focuses on bilingualism and multilingualism in children and adults, with particular emphasis on linguistic minorities, including heritage speakers, bilinguals, and Indigenous populations, as well as the interface between language policy and social inequality. She is co-editor of *Linguistic Approaches to Bilingualism*.

Chih-Cheng Lin is an Assistant Professor at Shih Chien University in Taipei, with academic interests in International Business, Marketing, EMI education, sustainability, and human–AI collaboration. Drawing on extensive international experience in Europe, his teaching and research emphasize cross-cultural learning, global business practice, ESG, and innovation management. He has contributed to teaching innovation through CDIO, COIL, and EMI initiatives, and actively mentors students in international academic and professional development.

Zi-Xu Liu is a university student and serves as Professor Lu's research assistant, supporting academic projects in the Department of Foreign Languages and Literature at National Cheng Kung University.

Hui-Chuan Lu is Professor of Spanish at National Cheng Kung University. Her research interests include eye-tracking, corpus linguistics, comparative syntax, and Spanish linguistics, with a particular interest in the processing and acquisition of Spanish as a foreign language.

Linh Thuy Nguyen is a Lecturer at the International Education Center, home to transnational programs at Hanoi University, Viet Nam. Linh holds an MA in TESOL & FLT, granted by University of Canberra (Australia). Her research interests include TESOL, CLIL and

sociolinguistics. She can be contacted at linhthuy.work210@gmail.com

Tran Tan Phat has a PhD in English Language Teaching, and is now a PhD candidate in TESOL at Ho Chi Minh City Open University. He is a lecturer in the Faculty of Foreign Languages at Ho Chi Minh University of Banking, Vietnam. He has more than 20 years of experiences in teaching, management, and academic work. His areas of interest include EMI, CLIL, Language Assessment and Evaluation, Business English, and ELT.

Adriani Yulia Purwaningrum is an Assistant Professor at the Faculty of Languages, Arts, and Cultures, Universitas Negeri Yogyakarta, specializing in English Language Teaching (ELT). She holds a Doctoral Degree in Language Education Sciences from Universitas Negeri Yogyakarta, Indonesia. Her research and pedagogical interests center on integrated language skills instruction and the application of technology in language learning. She is committed to advancing evidence-based, inclusive pedagogical practices that ensure language education remains meaningful, relevant, and accessible across diverse and evolving learning contexts.

Shigenori Wakabayashi received his PhD from the University of Cambridge. His research interests lie in applied linguistics, particularly in modeling second language learners' mental representations of morphology and syntax within the framework of generative grammar and psycholinguistics. Professor Wakabayashi was a founding member of the Japan Second Language Association (J-SLA) and has served as its Secretary and President. As of April 2026, he is once again serving as President of the association.

Taimin Tammy Wu is currently an Assistant Professor in the Department of English Language, Literature and Linguistics at Providence University. Her research interests center on second language writing, particularly source-based writing and learner agency in the AI era. She teaches English Composition, Academic Writing,

Studies in Writing Instruction and TESOL-related courses at both the undergraduate and graduate levels.

Wei-Wei Zhao is an Associate Professor in the Department of Chinese International Education at Quanzhou Normal University, China. His research interests include Buddhist philosophy, Chinese philosophy, and second language acquisition. His recent research focuses on interdisciplinary studies of Buddhist cognition and second language acquisition, as well as Buddhist culture in Fujian and Taiwan.

LINGUISTIC RELATIVITY AND L2 ACQUISITION: CLASSIFIER SENSITIVITY IN INTERMEDIATE MANDARIN LEARNERS

Laura Friedrich
University of Tübingen

Tanja Kupisch
Lund University

ABSTRACT

While previous research in the domain of linguistic relativity has mainly focused on how language affects cognition in L1 speakers, this study examines how German L2 learners of Mandarin Chinese conceptualize classifiers. L1 and L2 Mandarin speakers, as well as L1 German speakers, rated the similarity of two pictures which either matched in Mandarin classifier, German gender, both, or neither. This design allowed for a direct comparison of potential classifier and gender effects, two domains well-researched in linguistic relativity. Neither the gender nor the classifier manipulation affected the similarity ratings in either group. However, within the L2 Mandarin group, a significant interaction between the classifier modification and a combined measure of Mandarin proficiency and exposure suggested that a classifier effect might emerge in intermediate L2 learners and diminish with increasing proficiency. These findings seem to indicate that classifier effects in L2 learners are not stable or automatic but reflect a period of heightened sensitivity to the semantic information encoded in classifiers during acquisition.

Keywords: classifiers, gender, linguistic relativity, Mandarin, German

1. INTRODUCTION

Classifier systems such as the one used in Mandarin Chinese are a domain in which effects of linguistic relativity can surface. Ideas underlying linguistic relativity date back to Humboldt (1836/1988), who viewed language and thought as inseparable, and to Boas (1938), who argued that language shapes how experience is categorized, often outside speakers' awareness. These views later became associated with the Sapir-Whorf Hypothesis, although Sapir and Whorf never formulated a joint theory, as Sapir described language as a guide to social reality, while Whorf emphasized that grammatical differences can lead to different worldviews (Lucy 2016). Brown (1976) distinguished a weak version of linguistic relativity (language influences thought), from a strong, deterministic version (language determines thought), which Sapir and Whorf did not endorse. Research has since examined linguistic relativity across domains such as color, space, time, number, classifiers, and gender. Across all areas, there is no evidence for the strong version of linguistic relativity (Wolff and Holmes 2011). While language can influence thought, these effects do not override preexisting cognitive categories and can disappear under verbal interference or through learning new conceptual frameworks (e.g., Boroditsky 2001). Many reported effects are attributed to thinking-for-speaking (Slobin 1996), whereby language directs attention during speech planning. Even if such effects are often transient and task-dependent ("trivial Whorfianism", Reines and Prinz 2009), a relevant question is whether they can also surface in languages acquired later in life.

Our study focuses on linguistic relativity effects in the use of classifiers and grammatical gender, asking whether and how language affects cognition in late second language learners (L2ers). To date, few studies have investigated whether effects arising in monolingually raised speakers of a language can also emerge in late learners of that same language, posing the question whether L2ers can adopt the thought patterns of their L2 in a way that will affect their mental categorization. Another particularity of our study is that the L2ers already have a classification system in their L1 that is based on grammatical gender. We employ a similarity rating task with picture stimuli of item pairs sharing

the same Mandarin classifier, the same German classifier, both, or neither, to directly compare the effect that classifier and gender might have on mental categorization. The task will be completed by three groups of speakers: native Mandarin speakers, native German speakers (without knowledge of Mandarin), and native German speakers learning Mandarin as an L2. We investigate whether a classifier effect surfaces, and if so, how this is modulated by proficiency and exposure in the L2 group. In what follows, we will provide some background on nominal classification systems and previous research on relativity effects concerning classifiers and gender, before presenting our study and results. Throughout, there will be a stronger focus on classifier systems, as relevant to the target L2.

2. BACKGROUND

2.1 Classifiers and Gender

The term *classifier* can be used either to refer to systems of linguistic categorization in general or to describe a more specific set of classifying elements (Grinevald 2015). We adopt the latter approach. Classifiers occur in various grammatical constructions, yielding different types such as numeral, noun, possessive, verbal, locative, and deictic classifiers (Aikhenvald 2019). Numeral classifiers, the most widespread type, are found in approximately 140 of the 400 languages surveyed in the World Atlas of Language Structures (Gil 2013). The present study focuses exclusively on numeral classifiers, which categorize nouns according to properties such as animacy, humanness, shape, or function (Croft 1994).

Mandarin is a numeral classifier language that requires the use of classifiers when expressing quantity as well as anaphoric or deictic reference (Kuo and Sera 2009). The use of classifiers is obligatory in noun phrases with numerals and demonstratives:

- (1) a. *這 書
zhè shū
DEM book
- b. 這 本 書
zhè běn shū
DEM CLF book
'this book'

The classifiers in a numeral classifier language can be divided into two types that behave differently from a semantic and sometimes syntactic point of view (Her et al. 2017). *Sortal classifiers* categorize referents based on inherent properties and are used together with count nouns (Her and Hsieh 2010). *Mensural classifiers* can occur with mass and count nouns, denoting quantity or measurement units (Aikhenvald 2000). Sortal classifiers highlight inherent characteristics of the referent noun. One and the same noun in Mandarin Chinese can occur with different classifiers depending on the context, as shown in (2) for *fish*: combination with 隻 (*zhī*) (classifier for animals) in (2a) emphasizes animacy, while combination with 條 (*tiáo*) (for long, thin, flexible objects) in (2b) highlights the fish's shape (Her et al. 2022). In (2c), *fish* is classified by *box*, resulting in 'three boxes of fish', which highlights information about the container in which the fish is stored rather than the inherent features of a fish or their number. The mensural classifier *box* can be used for any object stored in a box, irrespective of its features. It can easily be replaced by any other container or weight expression (e.g., *three kilograms of fish*).

- (2) a. 三 隻 魚
sān zhī yú
three CLF.ANIMAL fish
'three fish'
- b. 三 條 魚
sān tiáo yú
three CLF.LONG fish
'three fish'

- c. 三 箱 魚
 sān xiāng yú
 three MENS.BOX fish
 ‘three boxes of fish’

(examples taken from Her et al. 2022: 154)

In Mandarin, the most frequently used classifier is 個 (*gè*), which not only accompanies all nouns without any special classifier but can substitute for many specific classifiers in informal speech. Gao and Malt (2009) distinguish classifier categories according to the prototypicality of their members. Classifiers in the *Well-Defined* category are highly transparent, as all objects using the same classifier share common features. An example is 本 (*běn*), which classifies all sorts of books and is limited to bound objects like books. In the *Prototype* category, classifiers occur with one or several highly prototypical objects whose features are associated with the classifiers, but they also accompany objects that resemble this prototype less strongly. For instance, 條 (*tiáo*) is often associated with long, thin, and flexible items, so that nouns like ropes, belts, or snakes are prototypical. Items like towels, trousers, or a fish are somewhat distant from the prototype but still plausible due to a similar shape. However, members like ships, experience, or shorts lost the transparent connection to the prototype. Last, classifiers in the *Arbitrary* category do not share defining features or a prototype but seem to be composed arbitrarily. There might have been a connection historically, but speakers of Mandarin have no clear intuition about this. An example is 尊 (*zūn*), which is used for large guns and Buddha statues.*

Beyond classifiers, languages may categorize nouns via grammatical gender or noun class systems, which resemble classifier systems in function but differ in several respects (Dixon 1986). Crucially, in gender systems, each noun is assigned to a fixed category, and gender systems typically involve two or three categories related to sex or gender, whereas noun class systems may include up to 20 categories based on features such as animacy, shape, or function. Gender assignment is often arbitrary and

* Classifier languages have often been claimed to treat all nouns as mass (Kuo 2003), but this view is disputed (Allan 1977; Cheng and Sybesma 1999; Zhang 2013).

language-specific: for example, German distinguishes masculine, feminine, and neuter, and inanimate nouns can belong to any category. There are some phonological, morphological, and semantic cues to gender, but they merely reflect tendencies (Köpcke 1982).

2.2 Classifier Effects

Zhang and Schmitt (1998) showed that Mandarin speakers perceive objects sharing the same classifier as more similar than English speakers. In similarity-rating tasks, classifier-sharing pairs received higher ratings, with a stronger effect among Mandarin speakers. When listing common features of such object pairs, Mandarin speakers mentioned classifier-related features earlier. In recall tasks, they also showed greater clustering of classifier-sharing items, although this effect weakened after controlling for overall recall.

About a decade later, Saalbach and Imai (2007) questioned the robustness of classifier effects by comparing them with taxonomic and thematic relations. Across object-selection, similarity-rating, induction, and priming tasks with Mandarin and German speakers, thematic and taxonomic relations consistently outweighed classifier-based categorization. Classifier effects emerged only in certain tasks, especially similarity ratings, and they were weaker than other conceptual relations. German speakers also showed sensitivity to classifier relations, suggesting that such similarities are perceptible even without a classifier system. Imai and Saalbach (2010) replicated these findings and added Japanese speakers. While Mandarin speakers again showed classifier effects, Japanese speakers did not, likely due to the less frequent use of classifiers in Japanese. Extending this work to Mandarin- and German-speaking children, Imai et al. (2010) found classifier effects only in flexible classification tasks; more constrained tasks elicited taxonomic or thematic preferences regardless of language. Overall, classifier effects appeared task-dependent and subordinate to other conceptual relations.

Gao and Malt (2009) examined classifier effects in a recall task with Mandarin and English speakers using sentence blocks containing different classifier types (well-defined, prototype, arbitrary, mixed). Both groups recalled items with well-defined classifiers best. Mandarin speakers

showed greater clustering (i.e., grouping stimuli with the same classifiers together), and this increased significantly only when well-defined classifiers were explicitly mentioned, suggesting that classifier transparency is crucial.

Huang and Chen (2014) similarly found that Mandarin speakers clustered nouns taxonomically unless classifier labels were provided, in which case classifier-based clustering increased. They concluded that classifier categories function as ad hoc categories that require explicit activation.

Speed et al. (2021) used similarity-rating and sorting tasks with Mandarin and Dutch speakers. Objects sharing a classifier were consistently judged more similar than distractors, but the effect was equally strong in both groups, while taxonomic classifications remained preferred. The authors therefore argued that classifiers reflect conceptual organization rather than shaping mental categories.

Srinivasan (2010) investigated classifier effects in a speeded visual counting task with Mandarin speakers and English- and Russian-speaking controls. Participants counted target objects among distractors that sometimes shared the same classifier. All groups were slower on classifier-sharing trials, likely due to visual similarity, but Mandarin speakers showed significantly greater slowing, indicating classifier-based interference in an automated process.

Tsang and Chambers (2011) examined classifier processing in Cantonese speakers. In classifier-judgment tasks, participants strongly preferred grammatically compatible objects, even when they did not match the classifier's prototypical shape. Eye-tracking revealed increased fixations on objects grammatically compatible with a classifier, whereas shape-matching but grammatically incompatible objects attracted no such advantage. Shape-based effects emerged only under modified conditions that highlighted classifier semantics. Overall, classifier effects were relatively weak and highly task-dependent.

Huettig et al. (2010) also used eye-tracking to examine online processing in Mandarin speakers. When classifiers were explicitly presented, participants fixated objects belonging to the corresponding classifier category more than distractors, indicating that classifiers can guide visual attention. However, when classifiers were omitted and only

indirectly activated through nouns, no fixation advantage emerged. The authors concluded that classifier effects are most likely to emerge in tasks that encourage explicit use of linguistic categories rather than in automatic conceptual processing.

2.3 Classifier Effects in Late Bilinguals

Although classifier effects have been widely studied, research on bilinguals remains limited. Song (2009) examined Mandarin and English speakers using a shape-similarity task with picture displays containing classifier-matching and non-matching distractors. In a picture-only condition, neither group responded faster when shape-matching items shared a classifier with the target, suggesting reliance on perceptual shape rather than classifier information. Classifier-matching distractors slowed responses in both groups, indicating general interference rather than a classifier effect. When labels were added, responses were faster for classifier-matching items, suggesting a label-induced classifier effect. However, the absence of an English control group in this condition limited interpretation. English learners of Mandarin also showed a non-significant tendency toward faster responses to classifier-matching items. Song attributed the inconsistent findings to heterogeneous participant backgrounds and concluded that there was no clear evidence for classifier effects in bilinguals.

Tang (2011) compared early Mandarin–English bilinguals in Singapore with late bilinguals who learned English as an L2. Participants judged whether noun pairs were related, with pairs sharing a classifier, a functional/thematic relation, or neither. Functional/thematic pairs received the highest ratings in both groups, but only late bilinguals rated classifier-related pairs as more related than unrelated ones, suggesting a classifier effect. In a recall task, both groups remembered functional/thematic pairs better than classifier-related or unrelated pairs, which did not differ.

Zhu (2021) tested native Mandarin speakers, advanced Mandarin learners with Russian as their L1, and a control group without Mandarin knowledge. In a similarity-rating task, both native speakers and learners rated classifier-related pairs as more similar than unrelated pairs, whereas the control group did not. In a recall task, learners showed greater

classifier-based clustering when labels were provided, suggesting that advanced L2 speakers can use classifiers to organize memory when relevant linguistic cues are available.

Overall, most classifier-effect research has focused on Mandarin speakers, and evidence from online-processing tasks remains limited. In recall studies, classifier effects usually emerge only in clustering behavior, with speakers of classifier languages showing slightly greater classifier-based organization than speakers of non-classifier languages. Similarity-judgment tasks that require categorical choices tend to produce weaker effects than Likert-scale ratings, which allow more nuanced responses. Several studies also indicate that classifier effects are strengthened when linguistic labels accompany visual stimuli.

2.4 Gender Effects

The question of whether grammatical gender influences the conceptualization of inanimate objects has generated considerable interest. A meta-analysis by Samuel et al. (2019), covering 43 studies and 5,895 participants, found mixed evidence: 32% of studies supported a gender effect, 24% yielded mixed results, and 43% found none. Effects were most common in tasks that made gender salient, such as voice-assignment and sex-attribution tasks, suggesting that they may reflect task strategies rather than stable conceptual differences. Gender effects were also more likely for animate entities and in languages with two rather than three gender categories.

Vigliocco et al. (2005) found gender effects among Italian speakers in semantic similarity and substitution tasks, particularly for animals and verbal stimuli, whereas German and English speakers showed no comparable effects. The effects disappeared with picture stimuli, suggesting a reliance on verbal mediation. In German, effects emerged only when gender-marked articles were produced, indicating that overt grammatical marking rather than conceptual structure drives the effect. Kousta et al. (2008) replicated the substitution findings and showed that Italian speakers displayed gender effects only when tested in Italian, not in English, suggesting an influence of language mode.

For Arabic speakers (two-way gender system), Almutrafi (2015) found gender effects in a voice-assignment task but not in similarity ratings. Arabic and English speakers relied on semantic relatedness, not shared grammatical gender, when judging similarity. Similar conclusions were reached by Ramos and Roberson (2011) for Portuguese. Although gender influenced voice-assignment and similarity-rating tasks, no gender effects emerged in a “best-match” task using verbal and pictorial stimuli. The authors concluded that gender effects are strongest when gender is explicitly referenced and weaken when linguistic processing is not required. Sedlmeier et al. (2016) compared German and Tamil (language without gender) speakers using potency ratings, adjective production, and similarity judgments. Tamil speakers showed no gender effects, whereas German speakers did across all tasks. However, the effects were small and nearly disappeared in similarity ratings, which do not explicitly highlight gender.

Several studies have examined whether gender effects emerge more clearly in speeded tasks. Cubelli et al. (2011) found that Italian speakers, unlike English speakers, responded faster to picture pairs sharing grammatical gender. Because this effect disappeared under verbal interference while semantic effects remained, the authors argued that grammatical gender facilitates linguistic processing rather than shaping conceptual representations.

Boutonnet et al. (2012) compared English monolinguals and Spanish–English bilinguals in a picture-matching task. No gender effects appeared in reaction times, but ERP data revealed increased sensitivity to grammatical gender among bilinguals, suggesting unconscious access to gender information even when it is not task-relevant.

Using eye-tracking, Bobb and Mani (2013) showed that German-learning toddlers activated both grammatical gender and semantic information during lexical processing. Competition between gender and semantic cues influenced visual attention, indicating that gender information is available early in language development.

Overall, grammatical gender effects are most consistently found when gender is highly salient, verbally mediated, and tested in two-gender languages (Samuel et al. 2019). Although effects in offline tasks are often weak, automated measures such as ERP and eye-tracking studies provide

evidence that grammatical gender can be accessed unconsciously during language processing.

3. RESEARCH QUESTIONS

Although classifier effects emerge in native speakers under some conditions, the question of whether and under which conditions classifier effects also surface in late L2ers is unresolved: Song's (2009) study yielded no clear results; effects were absent in Tang's (2011) study of simultaneous Mandarin-English bilinguals in Singapore, while Zhu (2021) found effects but with highly advanced L2ers of Mandarin in China. Can the latter be reproduced with less advanced Mandarin learners and at which level of proficiency does it emerge? The L2ers in the present study speak a gender language (German) as their L1, which allows us to compare potential L1 and L2 noun classification effects in the same group of speakers. Although gender effects in German have rarely been demonstrated, they still need to be controlled for. To our knowledge, no previous study has combined two different domains of linguistic relativity to compare the effect size. Classifier and gender systems both divide nouns in a language into different categories, making it possible to incorporate both domains in one task.

Our study will investigate the following questions:

RQ1. Does German gender affect the perceived similarity of objects in native German speakers?

RQ2. Do Mandarin classifiers affect the perceived similarity of objects in native Mandarin speakers?

RQ3. Do Mandarin classifiers affect the perceived similarity of objects in native German speakers of Mandarin as an L2?

Based on previous findings, we hypothesize that the effect of grammatical gender in German will be weak or absent in the German group, because gender effects tend to be relatively non-salient in similarity rating tasks with three-gender languages (Samuel et al. 2019; Sedlmeier et al. 2016). By contrast, we expect Mandarin classifiers to influence

object categorization such that items sharing a classifier are perceived as more similar than those that do not (Imai and Saalbach 2010; Saalbach and Imai 2007; Zhang and Schmitt 1998; Zhu 2021). Regarding the third research question, we expect only highly advanced learners of Mandarin to exhibit a classifier effect comparable to native speakers, since previous effects have been shown primarily for near-native L2ers (Zhu 2021).

4. METHODS

4.1 Materials

We employed a similarity rating task, which has revealed classifier effects in previous studies, while gender effects typically only surface when gender is relevant to the task. Following the premise that effects of linguistic relativity should be detectable in nonlinguistic tasks, the task is carried out with pictures.[†] Since both German and Mandarin speakers performed the task, using pictures guarantees uniformity, while translations might have evoked different associations depending on the language.

The task involved rating the perceived similarity of item pairs on a 9-point Likert scale. As shown in Table 1, it included the following four conditions with eight stimuli pairs each: C+G+ (item pairs selecting the same Mandarin classifier and German gender), C+G- (item pairs selecting the same classifier but different genders), C-G+ (item pairs selecting different classifiers but the same gender), and C-G- (item pairs selecting different classifier and different genders). This 2×2 design yields four conditions with item pairs sharing the same Mandarin classifier, the same

[†] We are aware that Saalbach and Imai (2007) opted against using pictures, arguing that visual similarity between pictures would influence the task, and participants might name pictures unconsciously anyway. However, while we agree that seeing pictures might induce unconscious naming, verbal stimuli have the additional disadvantage of involving the processing of written forms. Thus, a picture-based task is comparatively more nonlinguistic than a verbal task. In addition, we know little about sound-grapheme relations in L2ers of Mandarin, which might further influence the results. A pretest was carried out to control for visual similarity.

German gender, both, or neither (for the full list of all stimuli, see <https://osf.io/43vsa/>).

Table 1. Experimental design of the similarity rating task with an example of an item pair for each condition.

	C+ (classifier-congruent)	C- (classifier incongruent)
G+ (gender-congruent)	same classifier + same gender: cloud (朵, feminine) – flower (朵, feminine)	different classifier + same gender: sweater (件, masculine) – mountain (座, masculine)
G- (gender-incongruent)	same classifier + different gender: face (張, neuter) – table (張, masculine)	different classifier + different gender: heart (顆, neuter) – bridge (座, feminine)

To prepare the list of stimuli, we first compiled a list of 80 nouns that were included in the HSK[‡] 1 to 6 word lists (HSK Academy 2022), ensuring that most Mandarin learners would be familiar with the stimuli, and the lexical frequency of all items was comparable. We then administered three pretests to select the final items.

4.2 Preparation

The choice of the final test items required three stages of pretesting (for a full documentation, see <https://osf.io/43vsa/>). The goal of Pretest I was to ensure that the classifiers mentioned in dictionaries were the ones contemporary speakers mostly used with the stimuli. Pretest I was carried out with five native Mandarin speakers from Taiwan. Participants were presented with the nouns in written form and asked to write down the classifier they most naturally use with each noun. They also rated each noun for prototypicality, indicating how well it fit the semantic characteristics of the associated classifier. For instance, the classifier 條

[‡] HSK is an international assessment for learners of Mandarin Chinese.

(*tiáo*) can be used with both *rope* and *shorts*; however, since *tiáo* typically characterizes long, thin, and flexible objects, its use with *rope* is more prototypical than with *shorts*. Of the 80 stimuli, 59 received the same classifier from at least 80% of the raters and were therefore included in the final stimuli list. In addition, five items with 60% agreement were retained in order to construct eight item pairs per condition. These lower-agreement items were only used in the conditions in which the paired nouns did not share a classifier, so that limited inter-speaker agreement on classifier choice was less consequential.

The goal of Pretests II and III was to control for semantic and visual similarity amongst the items, following Huettig et al. (2010) and Cubelli et al. (2011). The two pretests were carried out with seven German native speakers each, who rated either the semantic or visual similarity of the picture pairs, while ignoring other aspects of similarity. They confirmed that the four different conditions did not differ significantly from each other in terms of either visual or semantic similarity. Furthermore, it was ensured that the item pairs did not begin with the same sound, either in German or Mandarin. Sixteen additional stimuli pairs which were either visually similar (*guitar* and *violin*) or semantically similar (*honey* and *bee*) were included as control items. Each stimulus was represented by a colored drawing taken from the MultiPic Database (Duñabeitia et al. 2018).

4.3 Participants

Participants were recruited through personal contacts, flyers distributed on campus, and by contacting departments of Sinology at German and Austrian universities. The main group consisted of 35 Mandarin learners from Germany and Austria. Additionally, two control groups were included: 22 L1 Mandarin speakers from Taiwan and 24 L1 German speakers from Germany (henceforth ‘Mandarin controls’ and ‘German controls’).[§] At the time of testing, all participants were pursuing

[§] Some speakers in the Mandarin control group had grown up with both Mandarin and Taiwanese. Taiwanese belongs to the Southern Min variety of Chinese and is commonly spoken alongside Standard Mandarin in Taiwan. While classifier usage may differ in pronunciation and frequency, the varieties generally share the same core classifier system

a university degree or had already graduated. The L2 Mandarin group, for whom the study was more time-consuming, had the opportunity to enter a prize draw with a chance to win one of six vouchers, each valued at 15 euros.

Of the 81 participants, data from 17 individuals were excluded from the final analysis (two from the Mandarin group, two from the German group, and 13 from the L2 Mandarin group) for two reasons. First, some participants rated themselves as highly proficient in another language with either classifiers or gender. In this case, it would have been impossible to conclude whether a potential effect stems from German and Mandarin or any other language with a different gender or classifier system. Most participants indicated that they had beginner or intermediate proficiency in a gender or classifier language. They were not excluded from the sample, since it is nearly impossible to find participants without any knowledge of gender or classifier languages. Most German participants had been exposed to a Romance language, while many Taiwanese participants indicated basic knowledge of Korean or Japanese. Second, ten German learners of Mandarin were excluded from the analysis because they identified the classifiers in five or fewer of the 32 item pairs. This limited performance yielded too few valid data points for reliable statistical analysis. Moreover, their low accuracy suggests insufficient familiarity with the classifier system, making it unlikely that any meaningful classifier-related effects would be observed in their responses.

Thus, the final sample included 20 L1 Mandarin speakers, 22 L1 German speakers, and 22 L1 German L2 Mandarin speakers. Table 2 provides an overview of the participants' gender and age distribution across groups.

due to historical and typological similarities (Chappell 2001). Given that most native speakers of Mandarin also speak one or more regional varieties, this limitation was considered acceptable for the present study.

Table 2. Number, gender, and age of participants in each group.

Group	<i>N</i>	Gender(F/M)	Age M (SD)	Age Range
Mandarin	20	14/6	22.35 (2.56)	19 – 28
L2 Mandarin	22	15/7	25.68 (4.42)	20 – 40
German	22	13/9	22.59 (1.71)	20 – 25

The L2 Mandarin group included in the final sample had been learning Mandarin for an average of 5.55 years (range = 2-17 years). 14 of the 22 learners had spent an extended period abroad in a Mandarin-speaking environment. The average length of their stay was 14 months (range = 6-18 months). The participants' self-estimated level of Mandarin with an HSK level between 4 and 6, which corresponds to A2 to B2 on the CEFR (Fachverband Chinesisch e. V. 2010). For a more objective measure of proficiency, the LexCHI test (Wen et al. 2024) was administered. Scores range from -100% to 100%. The L2ers scored 22% on average (range = -20% to 80%). Wen et al. (2024) defined a score of approximately 70% as the threshold below which a participant is very unlikely to be a native Mandarin speaker. Thus, participants' Mandarin proficiency in this study ranged from low to near-native, with most speakers reaching a lower-intermediate level, which roughly aligns with their self-estimated proficiency.

4.4 Procedure

The study was conducted online using the platform SoSci Survey (Leiner 2024) on the server hosted by the University of Konstanz. Participants in the control groups completed the study in approximately 15 minutes, and the L2 Mandarin group in about 45 minutes. All participants completed a background questionnaire and then performed the similarity rating task.

They rated the 32 critical and 16 control item pairs presented in pseudo-randomized order. Instructions were given in German for the German control group and the L2 Mandarin speakers, and in Mandarin for the Mandarin control group. Ratings were made on a 9-point Likert scale

ranging from *no similarity at all* to *extremely high similarity*. An example of the task is shown in Figure 1.

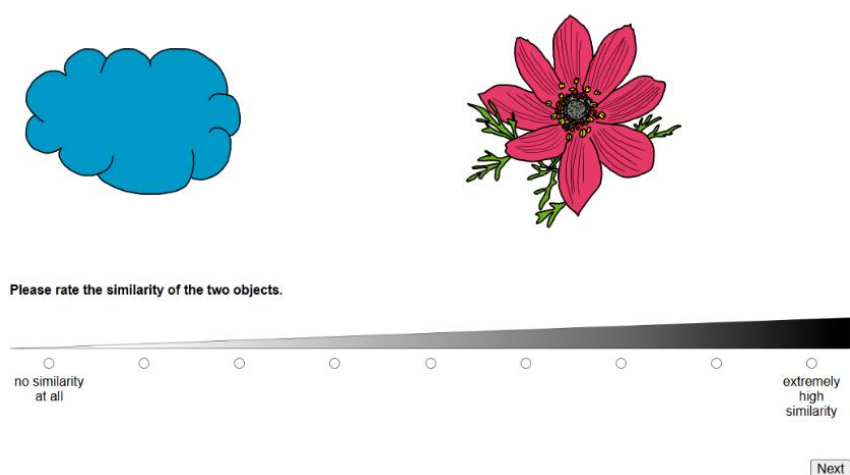


Figure 1. Example of the similarity rating task, asking participants to rate the similarity of the two stimuli on a 9-point Likert scale.

After completing the rating task, participants were asked which aspects had guided their ratings and whether they had recognized the purpose of the study. This question followed immediately after the similarity rating task to assess whether participants had noticed the gender or classifier manipulation during the rating task.

Following the rating task, the Mandarin learners named all critical picture stimuli in German to ensure they used the intended German nouns, suggesting they (potentially) activated the intended gender. Afterwards, they were asked to fill in the classifiers for the critical picture stimuli if they knew them (without providing the Mandarin name of the object) to check whether they had already acquired the classifiers for the relevant stimuli. At the end, the L2 Mandarin group completed the LexCHI (Wen et al. 2024), whose stimuli had also been incorporated into the SoSci survey. The procedure in the two control groups was comparatively shorter.

After the similarity rating task, the German group named the picture stimuli in German. The Mandarin controls named the objects in Mandarin together with a classifier. The questionnaire and all instructions can be found in the Supplementary Materials (<https://osf.io/43vsa/>).

4.5 Data Analysis

All ratings for which the participants had not provided the intended German gender, the intended Mandarin classifier, or both in the naming tasks were removed (16.7%; L2 Mandarin group: 39.5%, Mandarin group: 6.6%, German group: 3.0%). Although there is no guarantee that the items named in the naming task correspond to the ones the participants thought of during the similarity rating task, this method is the only way to check which name the picture stimuli spontaneously evoke in the participants, and consequently, exclude nontarget responses that do not fit the intended condition. Hence, this step ensured that statistical analyses were only performed on ratings for those items that subjects associated with the intended classifier or gender. If participants had written down a noun that was not the intended one, but still matched the intended gender or classifier, the rating was kept. In the incongruent conditions, which provided a baseline to which the items sharing the same gender or classifier were compared, ratings were kept as long as the gender or classifier of the pairs did not match. Thus, it was not required to provide the correct classifier. Instead, it was only relevant that the responses matched the conditions.

Statistical analyses were conducted using the R Statistical language (version 4.3.3; R Core Team 2024) on Windows 11 x64 (build 22631) with the packages *ordinal* (Christensen 2023), *emmeans* (Lenth 2024), *ggeffects* (Lüdtke 2024), and *factoextra* (Kassambara and Mundt 2020). Data wrangling and visualization were supported by the *tidyverse* (Wickham et al. 2019) package. To analyze the ordinal outcome variable (similarity ratings on a 9-point Likert scale), we fitted a cumulative link mixed model (CLMM). Previous studies with similar tasks (Saalbach and Imai 2007; Speed et al. 2021) had z-transformed the ratings within each participant to account for individual differences in how the scales had been used and then conducted a mixed analysis of variance (ANOVA).

However, this approach assumes normally distributed data. We tested the assumption of normality for the raw similarity ratings using the Shapiro-Wilk test. Results showed a significant deviation from normality ($W = 0.83$, $p < .001$), indicating that the data are not normally distributed. As the assumption of normality was violated, all analyses were conducted using a CLMM with a logit link function, which does not require normally distributed outcome variables.

5. RESULTS

5.1 Control Items

We first tested whether the conditions differed as expected, with control item pairs featuring high visual or semantic similarity hypothesized to receive the highest ratings. The initial CLMM included an interaction between Group and Condition. Participants and Item Pairs were included as random effects. A likelihood ratio test comparing this model to one with only main effects of Condition and Group showed that the interaction significantly improved model fit ($\chi^2 = 33.84$, $p < .001$). Post-hoc analyses using *emmeans* with Bonferroni correction for multiple comparisons revealed that, within each group, only the Control condition differed significantly from all other conditions ($p < .001$). Within the Control condition, the Mandarin and German groups differed significantly ($p = .006$). Significant differences between the Mandarin and L2 Mandarin groups were observed in the C+G+ condition ($p = .017$) and the C+G- condition ($p = .011$). All other pairwise contrasts were not significant. Figure 2 illustrates the model output.

5.2 Test Items

All further analyses focused only on critical item pairs. The four experimental conditions (C+G+, C+G-, C-G+, C-G-) were recoded into two binary variables to reflect the underlying 2×2 factorial design: Classifier and Gender, each coded as 0 (not shared) or 1 (shared), to examine the effect of shared classifier or gender independently. The initial

model included a three-way interaction between Gender, Classifier, and Group, since we expected an influence of German gender on the German group, an influence of Mandarin classifier on the Mandarin group, but an influence of both factors on the L2 Mandarin group. This model was compared to a model including two two-way interactions between Classifier and Group and between Gender and Group. A likelihood ratio test showed no significant improvement in fit with the three-way interaction ($\chi^2 = 3.79$, $p = .286$). Thus, the simpler model was retained. Adding another interaction between Classifier and Prototypicality (the score resulting from pretest I to examine how closely the objects matched the prototypical features of the classifier) did not improve the model fit either ($\chi^2 = 2.00$, $p = .368$). For this reason, the CLMM with the interactions between Classifier and Group and between Gender and Group, including random intercepts for Participant and Item Pairs, was considered for the analysis.

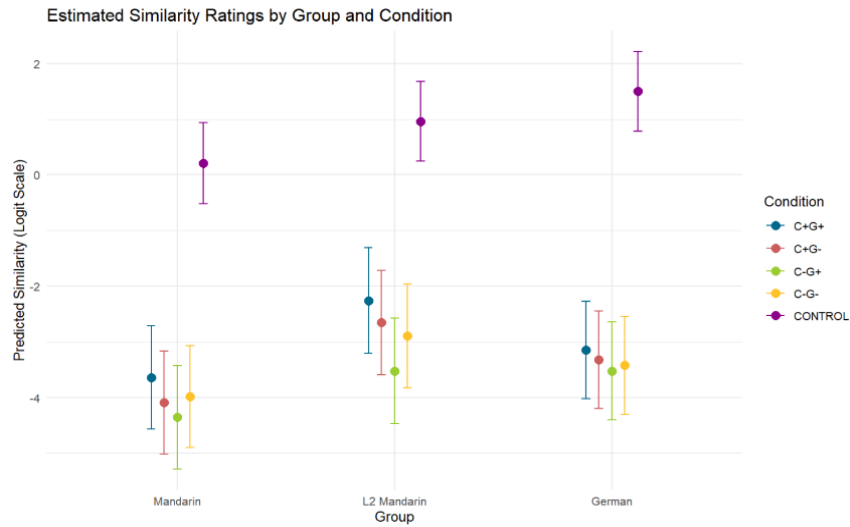


Figure 2. Estimated similarity ratings (logit scale) by Group and Condition, based on a CLMM. Higher logits reflect a greater likelihood of higher similarity ratings.

None of the interaction terms were significant (all $p > .230$), suggesting that the effects of shared classifier and gender did not differ between groups. Relative to the L2 Mandarin group, neither shared gender ($\beta = -0.27$, $p = .548$) nor shared classifier ($\beta = 0.65$, $p = .147$) was a significant predictor of similarity ratings. The L2 Mandarin group gave overall significantly higher similarity ratings than the Mandarin group ($\beta = -1.27$, $p = .040$), but not than the German group ($\beta = -0.41$, $p = .492$). Ratings from the German and Mandarin control groups also did not differ significantly from each other ($\beta = -0.86$, $p = .149$). Thus, Mandarin participants showed the lowest similarity ratings, L2ers the highest, with German participants falling in between but not significantly different from either. Figure 3 illustrates the group differences in similarity ratings by Classifier (left) and Gender (right), i.e., comparing items sharing the same classifier/gender (1) with those not sharing a classifier/gender (0).

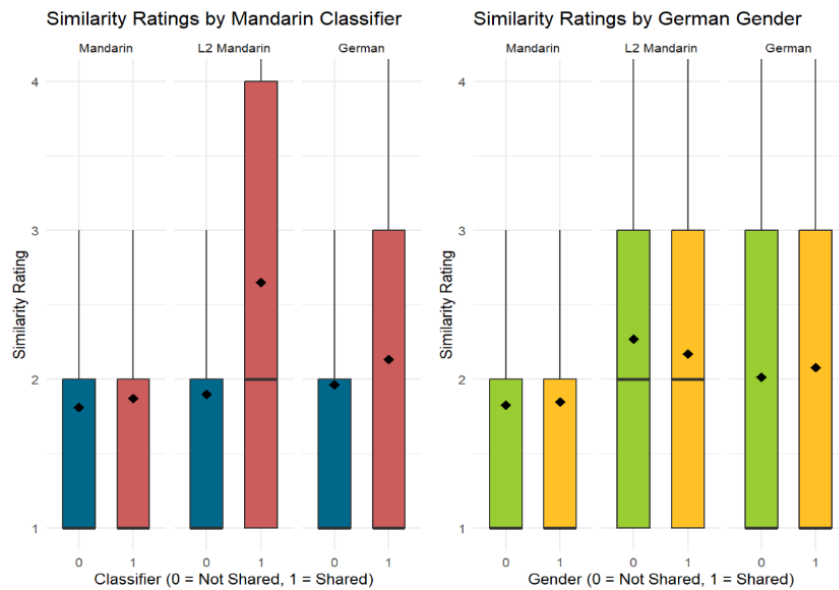


Figure 3. Distribution of similarity ratings by Group for both conditions: Mandarin Classifier (left) and German Gender (right). Ratings are shown on the original scale.

Lastly, we focused on the L2 Mandarin group only, who showed the greatest discrepancy in ratings between item pairs with incongruent and congruent classifiers. We wanted to investigate whether factors such as proficiency and exposure affected the ratings, as proposed in the third research question. A CLMM including an interaction between Classifier and any of the variables measuring proficiency or exposure did not show any significant effects. Main effects are disregarded here since they describe the effect of the variables on all similarity ratings. The interpretation of how proficiency and exposure influence learners' mental categorization of classifiers only becomes meaningful in an interaction with Classifier. A model including only Classifier as a predictor for the similarity ratings of the Mandarin learners did not reach significance ($p = .065$).

Since one of the variables did not seem to have a strong enough input and due to the high correlation between many of the factors, we carried out another exploratory analysis by scaling all predictors and implementing a Principal Component Analysis (PCA). The first component (PC1) seemed to be a good measure of both exposure and proficiency since it was strongly positively influenced by whether L2 learners had been abroad in a Mandarin-speaking environment, the duration of the stay, their self-estimated proficiency, the LexCHI as a measure of proficiency, and the number of pairs for which they had correctly named the classifiers. The duration of learning Mandarin, as well as the measures of active and passive exposure contributed less strongly, but positively, to the component.** Whether learners had identified the study goal was the only variable that negatively contributed to PC1. This factor is not necessarily related to the L2ers' proficiency or exposure, so we do not expect it to influence the interpretation of PC1.

We then included the interaction between Classifier and PC1 in the CLMM, again with Participants and Item Pairs as random intercepts. A likelihood ratio test showed no significant improvement in fit when adding interactions between Classifier and Study Goal or between Classifier and Prototypicality. The simpler model was kept, showing a significant

** For the full analysis see: <https://osf.io/43vsa/>

interaction between Classifier and PC1 ($p = .019$). Post-hoc analyses using *emmeans* with Bonferroni correction for multiple comparisons revealed that participants with low ($-1\ SD$; $p = .006$) and average ($M = 0$; $p = .016$) levels of PC1 rated pairs with incongruent and congruent classifiers significantly differently. At high ($+1\ SD$; $p = .061$) levels of PC1, this difference was no longer statistically significant. This reflects the opposite pattern as expected, namely that the classifier effect decreases with higher levels of proficiency and exposure, as can be seen in Figure 4.

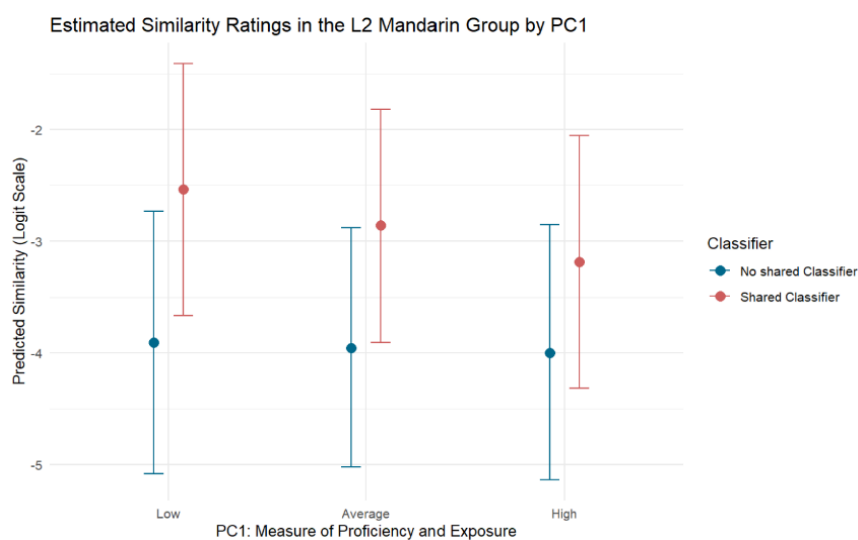


Figure 4. Estimated similarity ratings (logit scale) within the L2 Mandarin group, shown across levels of PC1 (a combined measure of proficiency and exposure). Higher logits indicate greater likelihood of higher similarity ratings.

6. DISCUSSION

Our study examined whether grammatical gender in German and classifiers in Mandarin influence the perceived similarity of objects in

native speakers of German and Mandarin, respectively, and whether Mandarin classifiers also affect similarity judgments in German learners of Mandarin as an L2. This was tested using similarity ratings that compared object pairs sharing the same gender or classifier with pairs that did not share these features.

6.1 Gender and Classifier Effects in L1 German and Mandarin Speakers

No effects of grammatical gender were observed (Figure 3), as predicted. This aligns with findings from Sedlmeier et al. (2016) for German, Almutrafi (2015) for Arabic, and Vigliocco et al. (2005) for German in offline similarity tasks. In contrast, gender effects have been reported when gender information is made salient; for example, with verbal stimuli (Ramos and Roberson 2011), in speech error tasks with gender-marked articles (Vigliocco et al. 2005), and in online measures such as eye-tracking, even in German (Bobb and Mani 2013). In the present study, the use of nonverbal stimuli, offline ratings, and a three-gender system likely reduced the salience of gender. This interpretation is supported by the debriefings: participants neither noticed the gender manipulation nor reported considering gender in their judgments. Together, this suggests that gender effects are least likely to emerge when gender marking is not perceptually or linguistically salient and when less sensitive offline methods are used.

Somewhat surprisingly, no significant classifier effects were observed either. This contrasts with earlier similarity-rating studies reporting classifier effects (e.g., Saalbach and Imai 2007; Imai and Saalbach 2010; Zhang and Schmitt 1998; Zhu 2021), but aligns with the more recent null findings of Speed et al. (2021).

One possible explanation for the absence of effects in our study compared to previous ones concerns stimulus selection. Classifier assignment in Mandarin is variable across regions, contexts, and speakers (Erbaugh 2013), and in previous studies nouns have not always been paired with the same different classifiers (e.g., *chair* with 把 (*bǎ*) vs. 張 (*zhāng*); *stamp* with 張 (*zhāng*) vs. 枚 (*méi*)), limiting comparability. In contrast to some of the previous studies, classifier assignments for the

chosen stimuli in our study were validated by native speakers and we retained only items with at least 80% agreement among raters. In addition, post-task naming ensured that only items named with the intended classifier were included in the analysis. Nevertheless, even validated items may not reliably activate the intended classifier associations, especially in the absence of context. Future work should therefore include transparent validation procedures and post-task naming checks.

A second possible explanation for the difference between our results and previous findings is related to prototypicality. In our study, prototypicality scores were based on five native-speaker ratings and showed limited variance after averaging across object pairs. This restricted range likely reduced sensitivity to prototypicality effects, which was not ideal. Future studies should systematically manipulate prototypicality across conditions to test its role more directly.

A third possible explanation for the absence of a classifier effect in our study concerns participants' linguistic environments. Earlier studies often tested predominantly monolingual speakers, whereas the present study, like Speed et al. (2021), involved multilingual participants with substantial exposure to English and other languages. This may attenuate language-specific conceptual effects. Supporting this idea, Kuo and Sera (2009) found stronger classifier-related preferences in Mandarin speakers in Taiwan than in those living in the U.S. Although we induced a Mandarin language mode through Mandarin instructions, participants' broader multilingual experience may have weakened classifier effects. Future research should therefore report and systematically vary L2 proficiency and language environment.

Importantly, our results differ from Speed et al. (2021) in that we did not observe classifier-congruent pairs to be rated as more similar, not even across groups. This challenges the proposal that classifiers enhance sensitivity to pre-existing semantic similarities (Saalbach and Imai 2007). We suggest that this discrepancy might be task-driven: most previous studies used multi-item comparison paradigms that encourage fine-grained differentiation, whereas our pairwise design likely promoted uniformly low similarity ratings. However, we would like to argue that if classifier-based similarity reflected a robust conceptual relation, it should also emerge in simple pairwise judgments. Its absence here, therefore,

suggests that classifier effects are weak, fragile, and highly task dependent, consistent with a weaker version of linguistic relativity.

6.2 Comparing Gender and Classifier Effects

Because no effect of either gender or classifiers emerged in the overall data, their effect sizes cannot be directly compared. However, the descriptive patterns suggest that classifiers may be more prone to inducing similarity effects than gender, which would be consistent with the fact that in the relevant research classifier effects have been reported more frequently than gender effects (e.g., Saalbach and Imai 2007). This is especially evident in the L2 Mandarin group, where no participant noticed the gender manipulation, while eight participants noticed the classifier manipulation.

Although both systems categorize nouns, they differ substantially. Classifier use is variable across contexts, dialects, and registers, and specific classifiers are often replaced by the general classifier 個 (*gè*) in informal speech (Erbaugh 2013). Classifiers are also acquired relatively late (Hao et al. 2021) and are infrequent in everyday discourse, occurring mainly in quantified or demonstrative contexts. Grammatical gender, by contrast, is acquired early (Kupisch et al. 2026; Szagun et al. 2007), it is relatively stable, and it is required in a wider range of syntactic environments. Despite this, gender appears to exert even weaker cognitive effects than classifiers, possibly because it encodes limited semantic information, whereas classifiers encode shape, size, and function. As a result, classifier categories group nouns with greater semantic overlap than grammatical gender does, which may explain why classifier-based effects, when they occur, are more robust than gender-based ones.

6.3 Classifier Effect in Intermediate-Level L2 Mandarin Speakers?

We expected the strongest classifier effects in native Mandarin speakers and weaker effects in L2 learners. Contrary to expectations, the largest (though non-significant) difference between classifier-congruent and incongruent pairs was observed in the L2 group ($p = .065$). Moreover,

it occurred in the group we deemed intermediate (low proficiency speakers were excluded due to insufficient classifier knowledge).^{††}

In an exploratory analysis, a composite measure combining residence in a Mandarin-speaking environment, duration of study, assessed and self-rated proficiency, classifier naming accuracy, duration of studying Mandarin, as well as active and passive exposure interacted with classifier ratings such that the effect emerged mainly for learners with low to average scores but not for learners with higher scores. Since we had excluded learners with low proficiency, these learners are presumably at an intermediate stage of proficiency. These, presumably intermediate, learners are likely to be in the process of developing a more automatic and flexible use of the language and are increasingly exposed to Mandarin outside of the classroom. It seems plausible that, at this stage, learners stop over-relying on the general classifier *gè* and develop first awareness of a systematic use of classifiers.^{‡‡} Based on our data, we tentatively propose that during intermediate stages, learners attend closely to classifier semantics, which surfaces in increased sensitivity to classifier semantics, before classifier use becomes automatized.

This interpretation is supported by processing studies showing that L2 learners rely more on semantic than grammatical cues when interpreting classifiers, whereas native speakers show little sensitivity to semantic competitors (Grüter et al. 2020; Tsang and Chambers 2011). Similarly, in

^{††} An anonymous reviewer raised the concern that, because we excluded the ten participants with very limited classifier knowledge, the observed “intermediate learner effect” (discussed below) might reflect a biased sample rather than a general pattern. Specifically, this raises the concern that the remaining participants may have been unusually sensitive to classifier distinctions. However, classifier knowledge is a prerequisite for making classifier-based similarity judgments, making it inappropriate to retain participants who lacked sufficient knowledge of the relevant classifiers. Furthermore, the sample also included more proficient learners, who demonstrated adequate classifier knowledge but did not exhibit classifier effects in the similarity-rating task, resembling the pattern observed among native speakers. This suggests that the intermediate learner effect cannot be attributed solely to the exclusion of participants with insufficient classifier knowledge.

^{‡‡} Even monolingual children acquire classifiers slowly, with heavy reliance on *gè* well into childhood (Hao et al. 2021). L2ers show a similar trajectory, moving from omission to overgeneralization and only later to target-like use (Gao 2010; Hansen and Chen 2001; Lau and Grüter 2015).

the present study, L2ers may have used their explicit knowledge of semantic cues for different classifiers, which are introduced in Mandarin textbooks and classes to facilitate acquisition, to guide similarity judgments, whereas native speakers and advanced learners processed classifiers primarily as grammatical markers. Because classifiers were not overtly present in the task, they may not have been salient for these groups. Overall, the results seem to suggest that classifier effects may be strongest at intermediate stages of acquisition, when learners are actively attending to classifier semantics, and diminish as classifier use becomes grammaticalized and automatic.

We acknowledge that, while plausible, this interpretation of our exploratory analysis warrants caution, since, overall, the classifier effect in the L2 group did not reach significance, the sample size was relatively small, and PC1 combines multiple partially overlapping constructs, complicating its interpretability. Future studies are therefore necessary to confirm our suggestion of heightened sensitivity to classifiers at an intermediate stage of L2 acquisition.

Interestingly, none of the native speakers identified the study goal, whereas eight of the 22 L2ers did. Although not statistically significant, there was a trend for goal-aware learners to show stronger classifier effects, suggesting that partial awareness may have amplified sensitivity to classifier information—a factor rarely controlled for in previous research. Moreover, all participants in the L2 group knew they were recruited for this study because of their knowledge of Mandarin. Thus, although we aimed to tap an implicit effect, the detected classifier sensitivity is likely, at least in part, a conscious effect. What we observed then seems to be best described as a heightened sensitivity to classifiers resulting from increased metalinguistic awareness of classifier categories during an intermediate stage of L2 acquisition, rather than as evidence of implicit conceptual restructuring. While this interpretation is less consistent with accounts of linguistic relativity that assume stable and automatic conceptual effects of language, it remains compatible with weaker approaches according to which linguistic categories can temporarily direct attention to semantic features encoded in the classifier system.

Overall, the development identified in the present work contrasts with previous studies which found a classifier effect in native-like Mandarin

L2ers (e.g., Zhu 2021). Our literature overview has shown task effects, and we acknowledge that another experimental design might have led to a different outcome. Nevertheless, similarity ratings remain a valid method because they have yielded valuable results in previous research.

6.4 Limitations

The study revealed some methodological challenges. The main difficulties were related to finding a large enough sample of L1 German L2 Mandarin speakers proficient enough to know classifiers and without knowledge of additional classifier or gender languages.

Moreover, classifier knowledge in the L2 Mandarin group was more limited than expected, which led to the exclusion of a substantial portion of the data from the analysis. In this study, only a combined measure of proficiency and exposure predicted the presence of a classifier effect. We acknowledge that subtle Whorfian effects may require substantially larger samples and could provide deeper insight into which specific variables are most relevant for the emergence of such effects.

A second methodological concern is the design, specifically the inclusion of control items and the use of pairwise contrasts, which may both have contributed to the absence of a classifier effect. As desired, the control pairs received significantly higher similarity ratings than any of the critical conditions. However, it is possible that the contrast with highly related control items made participants rate many critical pairs as minimally similar. While previous work (Saalbach and Imai 2007; Zhu 2021) has shown that classifier effects can emerge even when stronger semantic relations are present, it is still possible that participants rely on classifier information only when other, more salient similarities are absent. However, control items were necessary to ensure the use of the full scale.^{§§} Interestingly, the control items were rated significantly higher by the German than by the Mandarin group, suggesting group differences in rating behavior. Mandarin speakers used the lowest scale point more often,

^{§§} Additional analyses excluding participants who used only the lowest rating in more than 90% of the critical items, or excluding lowest ratings altogether, did not reveal classifier or gender effects. (see full analysis: <https://osf.io/43vsa/>). The absence of effects therefore cannot be attributed to floor effects.

possibly reflecting different response styles or expectations regarding similarity. Cultural factors or sample size limitations may have contributed to this pattern and warrant further investigation. Overall, effects of linguistic relativity appear to be highly task-specific and can often be eliminated by relatively minor changes in task design. Moreover, behavioral data alone may be insufficient to capture subtle online processing effects and should ideally be complemented by processing-based measures.

In terms of future prospects, it has become increasingly difficult to examine the effects of a single language in isolation, as most speakers today acquire multiple languages either from birth or later in life. This makes it problematic to attribute cognitive effects to one language alone. In the present study, many participants had knowledge of additional gender or classifier languages. Potential confounds cannot be ruled out, as it is impossible to control which linguistic features participants draw on in a task such as a similarity rating. Future research should therefore document participants' linguistic backgrounds and the language context of the experiment in as much detail as possible.

Finally, most existing studies focus on Mandarin, while other classifier languages remain largely underrepresented. The absence of an effect for Japanese speakers in Imai and Saalbach (2010), compared to Mandarin speakers, suggests that classifiers may influence cognition differently across languages. A broader cross-linguistic perspective is therefore needed. Such an approach has already proven informative in research on grammatical gender, where the number of gender categories has been shown to modulate the likelihood of observing an effect (see Samuel et al. 2019).

7. CONCLUSION

The present study did not find effects of classifiers or grammatical gender on cognition in either native Mandarin or German speakers, corroborating previous findings that such effects are highly task-dependent and may not surface when the relevant linguistic distinctions are not salient in the task. Interestingly, however, our exploratory analysis

points to a classifier effect in intermediate learners of Mandarin, who are still acquiring the system and have not yet fully automatized it. This may indicate heightened sensitivity to classifier-related semantic features at intermediate proficiency levels, before classifier use becomes routinized at more advanced stages. Future research comparing learners across proficiency levels with larger samples could provide insight into whether and how linguistic structure interacts with cognition over the course of acquisition.

REFERENCES

- Aikhenvald, Alexandra Y. 2000. *Classifiers: A typology of noun categorization devices*. Oxford: Oxford University Press.
- Aikhenvald, Alexandra Y. 2019. Noun categorization devices: A cross-linguistic perspective. In *Genders and classifiers: A cross-linguistic typology*, eds. Alexandra Y. Aikhenvald and Elena I. Mihás, pp.1-29. Oxford: Oxford University Press.
- Allan, Keith. 1977. Classifiers. *Language* 53(2):285-311.
<https://doi.org/10.1353/lan.1977.0043>
- Almutrafi, Fatimah. 2015. *Language and cognition: Effects of grammatical gender on the categorisation of objects*. Newcastle: University of Newcastle dissertation.
<http://hdl.handle.net/10443/3047>
- Boas, Franz. 1938. Language. In *General Anthropology*, ed. Franz Boas, pp.124-145. New York: D. C. Heath and Company.
- Bobb, Susan C., and Nivedita Mani. 2013. Categorizing with gender: Does implicit grammatical gender affect semantic processing in 24-month-old toddlers? *Journal of Experimental Child Psychology* 115(2):297-308.
<https://doi.org/10.1016/j.jecp.2013.02.006>
- Boroditsky, Lera. 2001. Does language shape thought?: Mandarin and English speakers' conceptions of time. *Cognitive Psychology* 43(1):1-22.
<https://doi.org/10.1006/cogp.2001.0748>
- Boutonnet, Bastien, Panos Athanasopoulos, and Guillaume Thierry. 2012. Unconscious effects of grammatical gender during object categorisation. *Brain Research* 1479:72-79. <https://doi.org/10.1016/j.brainres.2012.08.044>
- Brown, Roger. 1976. Reference: In memorial tribute to Eric Lenneberg. *Cognition* 4(2):125-153. [https://doi.org/10.1016/0010-0277\(76\)90001-9](https://doi.org/10.1016/0010-0277(76)90001-9)
- Chappell, Hilary. 2001. Language contact and areal diffusion in Sinitic languages. In *Areal diffusion and genetic inheritance: Problems in comparative linguistics*, eds. Alexandra Y. Aikhenvald and Robert M. W. Dixon, pp.328-357. Oxford: Oxford University Press.
- Cheng, Lisa L.-S., and Rint Sybesma. 1999. Bare and not-so-bare nouns and the structure of NP. *Linguistic Inquiry* 30(4):509-542. <https://doi.org/10.1162/002438999554192>
- Christensen, Rune H. B. 2023. *ordinal: Regression models for ordinal data*. R package version 2023.12-4. CRAN. <https://CRAN.R-project.org/package=ordinal>
- Croft, William. 1994. Semantic universals in classifier systems. *Word* 45(2):145-171.
<https://doi.org/10.1080/00437956.1994.11435922>
- Cubelli, Roberto, Daniela Paolieri, Lorella Lotto, and Remo Job. 2011. The effect of grammatical gender on object categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37(2): 449-460. <https://doi.org/10.1037/a0021965>
- Dixon, Robert. M. W. 1986. Noun classes and noun classification in typological perspective. In *Noun classes and categorization: Proceedings of a symposium on categorization and noun classification, Eugene, Oregon, October 1983*, ed. Colette G. Craig, pp.105-112. Amsterdam: John Benjamins.

- Duñabeitia, Jon Andoni, Davide Crepaldi, Antje S. Meyer, Boris New, Christos Pliatsikas, Eva Smolka, and Marc Brysbaert. 2018. MultiPic: A standardized set of 750 drawings with norms for six European languages. *Quarterly Journal of Experimental Psychology* 71(4):808-816.
<https://doi.org/10.1080/17470218.2017.1310261>
- Erbaugh, Mary S. 2013. Classifier choices in discourse across the seven main Chinese dialects. In *Increased empiricism: Recent advances in Chinese linguistics*, ed. Zhuo Jing-Schmidt, pp.101-126. Amsterdam: John Benjamins.
<https://doi.org/10.1075/scld.2.05erb>
- Fachverband Chinesisch e. V. 2010. Statement of the Fachverband Chinesisch e.V. (Association of Chinese Teachers in German Speaking Countries) on the new HSK Chinese Proficiency Test. https://www.fachverband-chinesisch.de/fileadmin/user_upload/Der_Verband/Offizielle_Statements/FaCh_2010_HSK-Erklaerung_en.pdf
- Gao, Helena H. 2010. A study of Swedish speakers' learning of Chinese noun classifiers. *Nordic Journal of Linguistics* 33(2):197-229.
<https://doi.org/10.1017/S0332586510000156>
- Gao, Ming Y., and Barbara C. Malt. 2009. Mental representation and cognitive consequences of Chinese individual classifiers. *Language and Cognitive Processes* 24(7-8):1124-1179. <https://doi.org/10.1080/01690960802018323>
- Gil, David. 2013. Numeral classifiers. In *The world atlas of language structures online*, eds. Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology.
- Grinevald, Colette. 2015. Classifiers, Linguistics of. In *International Encyclopedia of the Social & Behavioral Sciences*, 2nd ed., ed. James D. Wright, pp.811-818. Oxford: Elsevier.
- Grüter, Theres, Elaine Lau, and Wenyi Ling. 2020. How classifiers facilitate predictive processing in L1 and L2 Chinese: The role of semantic and grammatical cues. *Language, Cognition and Neuroscience* 35(2):221-234.
<https://doi.org/10.1080/23273798.2019.1648840>
- Hansen, Lynne, and Yung-Lin Chen. 2001. What counts in the acquisition and attrition of numeral classifiers. *JALT Journal* 23(1):90-110.
<https://doi.org/10.37546/JALTJJ23.1-5>
- Hao, Ying, Lisa Bedore, Li Sheng, Peng Zhou, and Li Zheng. 2021. Exploring influential factors of shape classifier comprehension and production in Mandarin-speaking children. *First Language* 41(5):573-604.
<https://doi.org/10.1177/01427237211026435>
- Her, One-Soon, and Chen-Tien Hsieh. 2010. On the semantic distinction between classifiers and measure words in Chinese. *Language and Linguistics* 11(3):527-551.
- Her, One-Soon, Ying-Chun Chen, and Nai-Shing Yen. 2017. Mathematical values in the processing of Chinese numeral classifiers and measure words. *PLoS ONE* 12(9):e0185047. <https://doi.org/10.1371/journal.pone.0185047>

- Her, One-Soon, Harald Hammarström, and Marc Allasonnière-Tang. 2022. Defining numeral classifiers and identifying classifier languages of the world. *Linguistics Vanguard* 8(1):151-164. <https://doi.org/10.1515/lingvan-2022-0006>
- HSK Academy. 2022. *Welcome to HSK Academy!*. <https://hsk.academy/en/>
- Huang, Shuping, and Jenn-Yeu Chen. 2014. The effects of numeral classifiers and taxonomic categories on Chinese and English speakers' recall of nouns. *Journal of East Asian Linguistics* 23:27-42. <https://doi.org/10.1007/s10831-013-9108-0>
- Huetting, Falk, Jidong Chen, Melissa Bowerman, and Asifa Majid. 2010. Do language-specific categories shape conceptual processing? Mandarin classifier distinctions influence eye gaze behavior, but only during linguistic processing. *Journal of Cognition and Culture* 10(1-2):39-58. <https://doi.org/10.1163/156853710X497167>
- Humboldt, Wilhelm von. 1988. *On language: The diversity of human language-structure and its influence on the mental development of mankind*. Trans. Peter Heath. Cambridge: Cambridge University Press. (Original work published in 1836)
- Imai, Mutsumi, and Henrik Saalbach. 2010. Categories in mind and categories in language: Do classifier categories influence conceptual structures? In *Words and the mind: How words capture human experience*, eds. Barbara Malt and Philip Wolff, pp.138-164. Oxford: Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195311129.003.0008>
- Imai, Mutsumi, Henrik Saalbach, and Elsbeth Stern. 2010. Are Chinese and German children taxonomic, thematic, or shape biased? Influence of classifiers and cultural contexts. *Frontiers in Psychology* 1:194. <https://doi.org/10.3389/fpsyg.2010.00194>
- Kassambara, Alboukadel, and Fabian Mundt. 2020. *factoextra: Extract and visualize the results of multivariate data analyses*. R package version 1.0.7. CRAN. <https://CRAN.R-project.org/package=factoextra>
- Köpcke, Klaus-Michael. 1982. *Untersuchungen zum Genusssystem der deutschen Gegenwartssprache [Research on the gender system of contemporary German]*. Tübingen: Max Niemeyer. <https://doi.org/10.1515/9783111676562>
- Kousta, Stavroula-Thaleia, David P. Vinson, and Gabriella Vigliocco. 2008. Investigating linguistic relativity through bilingualism: The case of grammatical gender. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34(4):843-858. <https://doi.org/10.1037/0278-7393.34.4.843>
- Kuo, Jenny Yi-chun, and Maria D. Sera. 2009. Classifier effects on human categorization: The role of shape classifiers in Mandarin Chinese. *Journal of East Asian Linguistics* 18:1-19. <https://doi.org/10.1007/s10831-008-9036-6>
- Kuo, Yi-chun. 2003. *Shape salience in English and Chinese: Implications for the effects of language on cognition*. Minneapolis: University of Minnesota dissertation.
- Kupisch, Tanja, Cristina Flores, Silvina Montrul, and Sílvia Perpiñán. 2026. The acquisition of gender in Romance languages. In *The Oxford Handbook of Gender and Classifiers*, eds. Niels O. Schiller and Tanja Kupisch, pp.70-96. Oxford: Oxford University Press.
- Lau, Elaine, and Theres Grüter. 2015. Real-time processing of classifier information by L2 speakers of Chinese. In *BUCLD 39: Proceedings of the 39th annual Boston*

- University Conference on Language Development, vol. 2, eds. Elizabeth Grillo and Kyle Jepson, pp.311-323. Somerville, MA: Cascadilla Press.
- Leiner, Dominik J. 2024. *SoSci Survey*. Version 3.2.55. <https://ling-survey.uni-konstanz.de/sosci/admin/index.php>
- Lenth, Russell. V. 2024. *emmeans: Estimated marginal means, aka least-squares means*. R package version 1.10.1. CRAN. <https://cran.r-project.org/package=emmeans>
- Lucy, John A. 2016. Recent advances in the study of linguistic relativity in historical context: A critical assessment. *Language Learning* 66(3):487-515. <https://doi.org/10.1111/lang.12195>
- Lüdtke, Daniel. 2024. *ggeffects: Create tidy data frames of marginal effects for ggplot from model outputs*. R package version 2.3.0. CRAN. <https://CRAN.R-project.org/package=ggeffects>
- R Core Team. 2024. *R: A language and environment for statistical computing*. Version 4.3.3. Vienna: R Foundation for Statistical Computing. <https://www.R-project.org/>
- Ramos, Sara, and Debi Roberson. 2011. What constrains grammatical gender effects on semantic judgements? Evidence from Portuguese. *Journal of Cognitive Psychology* 23(1):102-111. <https://doi.org/10.1080/20445911.2011.466795>
- Reines, Maria F., and Jesse Prinz. 2009. Reviving Whorf: The return of linguistic relativity. *Philosophy Compass* 4(6):1022-1032. <https://doi.org/10.1111/j.1747-9991.2009.00260.x>
- Saalbach, Henrik, and Mutsumi Imai. 2007. Scope of linguistic influence: Does a classifier system alter object concepts? *Journal of Experimental Psychology: General* 136(3):485-501. <https://doi.org/10.1037/0096-3445.136.3.485>
- Samuel, Steven, Geoff Cole, and Madeline J. Eacott. 2019. Grammatical gender and linguistic relativity: A systematic review. *Psychonomic Bulletin & Review* 26(6):1767-1786. <https://doi.org/10.3758/s13423-019-01652-3>
- Sedlmeier, Peter, Arun Tipandjan, and Anastasia Jänchen. 2016. How persistent are grammatical gender effects? The case of German and Tamil. *Journal of Psycholinguistic Research* 45:317-336. <https://doi.org/10.1007/s10936-015-9350-x>
- Slobin, Dan I. 1996. From 'thought and language' to 'thinking for speaking'. In *Rethinking linguistic relativity*, eds. John J. Gumperz and Stephen C. Levinson, pp.70-96. Cambridge: Cambridge University Press.
- Song, Jiang. 2009. *The semantic structure of Chinese classifiers and its implications for linguistic relativity*. Honolulu: University of Hawai'i dissertation.
- Speed, Laura J., Jidong Chen, Falk Huettig, and Asifa Majid. 2021. Classifier categories reflect but do not affect conceptual organization. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 47(4):625-640. <https://doi.org/10.1037/xlm0000967>
- Srinivasan, Mahesh. 2010. Do classifiers predict differences in cognitive processing? A study of nominal classification in Mandarin Chinese. *Language and Cognition* 2(2):177-190. <https://doi.org/10.1515/langcog.2010.007>
- Szgun, Gisela, Barbara Stumper, Nina Sondag, and Melanie Franik. 2007. The acquisition of gender marking by young German-speaking children: Evidence for

- learning guided by phonological regularities. *Journal of Child Language* 34(3):445-471. <https://doi.org/10.1017/S0305000906007951>
- Tang, Huimin. 2011. *Language and thought: Classifier effect in early and late bilinguals*. Final year project, Nanyang Technological University. <https://dr.ntu.edu.sg/entities/publication/cfd8581c-679d-4df5-9f18-404161d16741>
- Tsang, Cara, and Craig G. Chambers. 2011. Appearances aren't everything: Shape classifiers and referential processing in Cantonese. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37(5):1065-1080. <https://doi.org/10.1037/a0023601>
- Vigliocco, Gabriella, David P. Vinson, Federica Paganelli, and Katharina Dworzynski. 2005. Grammatical gender effects on cognition: Implications for language learning and language use. *Journal of Experimental Psychology: General* 134(4):501-520. <https://doi.org/10.1037/0096-3445.134.4.501>
- Wen, Yun, Yicheng Qiu, Christine X. R. Leong, and Walter J. B. van Heuven. 2024. LexCHI: A quick lexical test for estimating language proficiency in Chinese. *Behavior Research Methods* 56(3):2333-2352. <https://doi.org/10.3758/s13428-023-02151-z>
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D. A. McGowan, Romain François, Garrett Golemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas L. Pedersen, Evan Miller, Stephan M. Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana P. Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. 2019. Welcome to the tidyverse. *Journal of Open Source Software* 4(43):1686. <https://doi.org/10.21105/joss.01686>
- Wolff, Phillip, and Kevin J. Holmes. 2011. Linguistic relativity. *Wiley Interdisciplinary Reviews: Cognitive Science* 2(3):253-265. <https://doi.org/10.1002/wcs.104>
- Zhang, Niina N. 2013. *Classifier structures in Mandarin Chinese*. Berlin: De Gruyter.
- Zhang, Shi, and Bernd Schmitt. 1998. Language-dependent classification: The mental representation of classifiers in cognition, memory, and ad evaluations. *Journal of Experimental Psychology: Applied* 4(4):375-385. <https://doi.org/10.1037/1076-898X.4.4.375>
- Zhu, Sanmao. 2021. An empirical study of the classifier effect in advanced L2 users of Chinese. *International Journal of Bilingualism* 25(1):151-167. <https://doi.org/10.1177/1367006920943301>

Laura Friedrich
Department of Linguistics
University of Tübingen
Tübingen, Germany
laura.friedrich@uni-tuebingen.de

Tanja Kupisch
Centre for Languages and Literature
Lund University
Lund, Sweden
tanja.kupisch@ling.lu.se

語言相對論與第二語言習得：
中階華語學習者對量詞的敏感性

Laura Friedrich
圖賓根大學

Tanja Kupisch
隆德大學

過去語言相對論 (linguistic relativity) 研究多著重語言如何影響母語 (L1) 使用者的認知，本研究則探討華語第二語言 (L2) 學習者如何表徵量詞系統。華語母語者、德語母語者及華語二語學習者針對不同量詞與語法性別配對之圖片進行相似度評定。結果顯示，量詞條件與語法性別條件在各組評分中均無顯著影響；然而在二語學習者組別中，量詞條件與一項結合華語能力及語言接觸程度之綜合指標間呈現顯著交互作用：量詞效應 (classifier effect) 出現於中階學習者，並隨能力提升而減弱。此結果顯示，二語學習者之量詞效應並非穩定或自動化歷程，而反映其習得過程中對量詞語意資訊之暫時性高度敏感期。

關鍵詞：量詞、語法性別、語言相對論、華語、德語

UNDERLYING SYNTAX OF LONG-DISTANCE WH- QUESTIONS IN L2 ENGLISH: INSIGHTS FROM PROSODY*

Tokitada Hangai and Shigenori Wakabayashi
Chuo University

ABSTRACT

This study investigates the underlying syntactic structures of long-distance *wh*-questions produced by Japanese learners of English (JLEs) using prosodic evidence. Previous studies have analysed learner productions based on surface word order alone, assuming that target-like forms reflect target-like syntax. Adopting the view that prosodic structure transparently reflects syntactic structure (Royer 2022), this study argues that Intonational Phrase (IntP) boundaries provide evidence for clause-level syntactic organisation. An Elicited Production Task revealed that target-like productions are not prosodically uniform: some were realised as a single IntP, others as two IntPs, suggesting structural differences undetectable from surface form alone. The *wh-wh* type similarly showed prosodic variation, indicating at least two distinct underlying derivations. The present findings indicate that prosodic analysis serves as a reliable diagnostic for investigating L2 grammar.

Keywords: *wh*-questions, English, Japanese, syntax–prosody interface, variation

* We thank two anonymous reviewers for their valuable comments. Our research was supported by a Shunju Scholarship Foundation to the first author, and a Chuo University Personal Research Grant to the second author.

1. INTRODUCTION

This study examines the underlying syntactic structures of second language (L2) grammar through the lens of prosody. A common assumption in Generative Second Language Acquisition (GenSLA) is that learners who produce grammatical sentences in the target language have acquired the same underlying syntax as native speakers. For example, if a learner produces a long-distance *wh*-question such as *Who do you think likes Mary?*, it is often assumed that the learner has acquired the target-like syntactic representation for this construction. However, this assumption is not necessarily justified. As Hawkins and Hattori (2006) points out, target-like behaviour in L2 learners does not always imply target-like underlying syntax. The central question of the present study, therefore, is whether learners who produce target-like surface forms necessarily employ the same syntactic representations as native speakers.

Previous studies (Kimura 2022; Schulz 2011; Wakabayashi and Okawara 2003; Yamane 2003) have examined JLEs' knowledge of long-distance *wh*-dependencies through elicited production tasks. These studies have shown that JLEs produce not only target-like forms but also non-target-like forms such as *What do you think who liked Mary?* (the *wh-wh* type) and *Do you think who liked Mary?* (the *do-wh* type). To account for these productions, previous research has appealed to factors such as Merge over Move (Wakabayashi and Okawara 2003), differences between English and Japanese *wh*-configurations and their derivations (Kimura 2022), and processing difficulty (Schulz 2011). While these studies have provided important insights into the nature of L2 syntax, they have largely relied on surface word order as the primary diagnostic. As a result, productions that are identical in surface form have typically been assigned a single underlying representation, leaving open the possibility that structurally distinct derivations may underlie what appears to be the same output.

The present study takes a different perspective. Drawing on theoretical work suggesting that prosody can provide evidence for syntactic structure (Royer 2022), it investigates whether prosodic

structure can reveal underlying syntactic variation in learner productions. The key idea is that prosodic phrasing is not arbitrary but reflects the syntactic constituency of the utterance: where a syntactic boundary is present, a prosodic boundary tends to emerge. Under this view, two utterances that are identical in word order but differ in prosodic structure may reflect genuinely different syntactic representations. Applied to L2 production, this means that prosody can potentially reveal hidden structural differences that surface form alone cannot detect.

This issue is especially relevant to long-distance *wh*-questions, because English and Japanese differ substantially in how *wh*-dependencies are realised. English requires overt *wh*-movement, whereas Japanese allows *wh*-phrases to remain *in situ*. This cross-linguistic difference has made long-distance *wh*-questions a central testing ground for research on L2 syntax. If prosody provides evidence for underlying syntactic structure, then learner productions that appear target-like on the surface may nevertheless involve non-target-like syntactic representations—a possibility that production studies based on word order alone cannot address.

The central claim of this study is that it provides us with an analytical tool to investigate learner syntax, and that in fact a type of learner grammar for appear-to-be target-like long-distance *wh*-questions is likely to differ from the native-speaker grammar. In particular, this study adopts the view that a single Intonational Phrase (IntP) corresponds to a single CP, whereas two IntPs suggest a structure containing two CPs (Royer 2022). If these correspondence holds, then target-like surface form does not necessarily imply target-like syntax. The learner data examined in this study support this claim: even target-like productions are not prosodically uniform, and the *wh-wh* type also exhibits prosodic variation. These findings suggest that learner productions are structurally more diverse than surface form alone would indicate.

The remainder of this paper is organised as follows. Section 2 reviews the relevant background on long-distance *wh*-questions, previous studies on their acquisition, and the syntax–prosody correspondence adopted in this study. Section 3 presents the research questions and predictions. Section 4 describes the method. Section 5 reports the results. Section 6

discusses the implications of these findings for the analysis of L2 grammar. Section 7 concludes the paper.

2. BACKGROUND

2.1 Long-distance *wh*-questions in English and Japanese

Long-distance *wh*-questions involve a dependency between a sentence-initial *wh*-phrase and a position inside an embedded clause, as shown in (1). (1a) illustrates a long-distance *wh*-question with object extraction, while (1b) illustrates one with subject extraction. In both cases, the sentence-initial *wh*-phrase is interpreted as the question word corresponding to the gap in the embedded clause, indicated as *t(race)*.

- (1) a. Who_i do you think Mary liked *t_i*?
 b. Who_i do you think *t_i* liked Mary?

In English, such dependencies are formed by overt *wh*-movement. This movement proceeds cyclically through the edge of the embedded clause rather than applying in a single step (Bošković 2007; Chomsky, 2000, 2001). In other words, the *wh*-phrase first moves to the edge of the embedded CP and then moves further to the edge of the matrix CP. This derivation is illustrated in (2).

- (2) a Step 1: Derivation of embedded clause
 $[_{CP}^{EMB} \text{ who } [_{uQ}] C_{[+Q]} [_{TP} \text{ Mary liked } \text{who} [_{uQ}]]]$
 ↑
 b Step 2: Derivation of matrix clause
 $[_{CP}^{MAT} \text{ Who } [_{uQ}] C_{[+Q]} \text{ do } [_{TP} \text{ you think } [_{CP}^{EMB} \text{ who } [_{uQ}] C_{[+Q]} [_{TP} \text{ Mary liked } \text{who} [_{uQ}]]]]$

In Step 1, the *wh*-phrase moves from its base position inside the embedded TP to the edge of the embedded CP. This movement is

necessary because the *wh*-phrase carries an uninterpretable feature [uQ] that must be valued in the course of the derivation, otherwise the derivation fails to converge (Chomsky 1995). Since the embedded C bears no [Q] feature, the [uQ] feature cannot be valued at Spec-CP^{EMB}, and the *wh*-phrase must therefore move further. As shown in Step 2, it moves to the matrix Spec-CP, where the [+Q] feature on the matrix C values and deletes [uQ], allowing the derivation to converge. The crucial point is that this movement is both obligatory and cyclic: the *wh*-phrase cannot reach the matrix CP in a single step but must pass through the edge of the embedded CP as an intermediate landing site.

Japanese differs from English in the optionality of *wh*-phrase movement, as shown in (3) and (4). (3a) illustrates a long-distance *wh*-question with the *wh*-phrase remaining *in situ*, and (3b–c) show counterparts in which the *wh*-phrase has been scrambled within the embedded clause and to the matrix clause, respectively. (4) shows the subject *wh*-question counterparts.

- (3) a. Anata-wa [Hanako-ga dare-o suki-da-ta to]
 you-TOP Hanako-NOM who-ACC like-COP-PST COMP
 omoi-masu ka.
 think-POL Q
- b. Anata-wa [dare-o_i Hanako-ga *t_i* suki-da-ta to]
 you-TOP who-ACC Hanako-NOM *t_i* like-COP-PST COMP
 omoi-masu ka.
 think-POL Q
- c. Dare-o anata-wa [Hanako-ga *t_i* suki-da-ta to]
 who-NOM you-TOP Hanako-NOM *t_i* like-COP-PST COMP
 omoi-masu ka.
 think-POL Q
 ‘Who do you think Hanako liked?’
- (4) a. Anata-wa [dare-ga Hanako-o suki-da-ta to]
 you-TOP who-NOM Hanako-ACC like-COP-PST COMP
 omoi-masu ka

think-POL Q

- b. Dare-ga anata-wa ti Hanako-o suki-da-ta to]
 who-NOM you-TOP ti Hanako-ACC like-COP-PST COMP
 omoi-masu ka
 think-POL Q
 ‘Who do you think liked Hanako?’

The scope of the *wh*-phrase is determined by its relation to the Q-particle *ka*. In (3a), for instance, the *wh*-phrase *dare-o* remains *in situ* inside the embedded clause, yet the Q-particle *ka* appears in the matrix clause, indicating that the *wh*-phrase takes matrix scope. While Japanese allows *wh*-phrases to remain *in situ* in their base-generated positions, it also allows them to be displaced to the matrix clause via scrambling (Saito 1992). Crucially, unlike English long-distance *wh*-movement, this displacement is optional: failure to scramble does not result in ungrammaticality. In other words, the grammaticality of Japanese long-distance *wh*-questions does not depend on overt movement of the *wh*-phrase to the matrix clause. This contrast suggests that Japanese long-distance dependencies differ fundamentally from English ones, where overt and successive-cyclic *wh*-movement is obligatory. The schematic derivation in (5) illustrates this point. In Step 1, the embedded clause is constructed with the *wh*-phrase in its base position. In Step 2, the matrix clause is built without forcing the *wh*-phrase to leave the embedded clause. In Step 3, the *wh*-phrase may optionally undergo scrambling to the matrix clause via embedded clause, but this operation is not required for the derivation to converge.

- (5) a. Step 1: Derivation of embedded clause
 [_{CP}^{EMB} [_{TP} NP-NOM *wh*-ACC V] COMP]
 b. Step 2: Derivation of main clause
 [_{CP}^{MAT} NP-TOP [_{CP}^{EMB} [_{TP} NP-NOM *wh*-ACC V] COMP] V Q]
 c. Step 3: Optional scrambling of *wh*-phrase to the matrix edge
 [_{CP}^{MAT} *wh*_i-ACC NP-TOP [_{CP}^{EMB} *wh*_i-ACC [_{TP} NP-NOM *t*_i V]

COMP] V Q]

The crucial contrast between the two languages, then, is not simply that both allow displacement of the *wh*-phrase, but that the status of that driving force is fundamentally different. In English, overt movement is obligatory and driven by the need to value an uninterpretable feature. In Japanese, by contrast, movement is optional and not required for convergence. This asymmetry is important for the present study because it raises the possibility that Japanese learners of English may produce target-like surface forms through different derivation from those of native English speakers (Hawkins and Hattori 2006; Kimura 2022; Kimura and Wakabayashi 2024). Since the L1 grammar does not require *wh*-movement, learners may not have fully internalised the obligatory and cyclic nature of the relevant operation in English, a possibility that cannot be detected from surface form alone.

2.2 Wh-scope-marking constructions

A further construction relevant to the present study is the *wh*-scope-marking construction. In this construction, a *wh*-element appears in the matrix clause while another *wh*-element appears in the embedded clause. Japanese provides examples of this pattern, such as those discussed by Fujiwara (2021). In these constructions, the embedded *wh*-phrase expresses the semantic content of the question, while the matrix *wh*-phrase functions as a scope marker that extends the interrogative scope to the matrix clause.

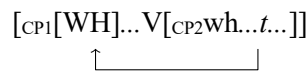
- (6) Ano-hito-wa [dare-ga senkyo-ni tousensi-soo ka]
 that-person-TOP who-NOM election-to win-EVID Q
 nante itte-ta kke?
 what_{PROP} say-PST Q
 ‘What did that person say: who would win the election?’
 (Fujiwara 2021:632)

In (6), the Q-particle *ka* appears in the embedded clause, indicating that the embedded *wh*-phrase *dare-ga* takes its scope within the embedded

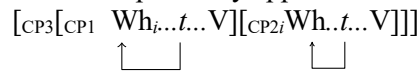
clause. Nevertheless, the sentence as a whole is interpreted as a question about a person, not as a yes-no question. This is because the matrix *wh*-phrase *nante* functions as a scope marker that extends the interrogative scope to the matrix clause. In other words, the embedded *wh*-word expresses what is being asked, while the matrix *wh*-word signals how far the scope of that question extends. This type of structure is known as a *wh*-scope-marking construction (Dayal 1993; Fujiwara 2021; Manetta 2010; McDaniel 1989).

Two major approaches have been proposed for the syntax of *wh*-scope marking: the direct dependency approach and the indirect dependency approach. Under the direct dependency approach (Manetta 2010; McDaniel 1989, a.o.), a direct syntactic relation holds between the matrix and embedded *wh*-elements. As shown in (7a), the embedded *wh*-phrase first moves to the edge of the embedded CP, and its feature then moves further to the Spec-CP of the matrix clause. On this account, *wh*-scope-marking constructions involve successive-cyclic movement, analogous to standard long-distance *wh*-movement, as shown in (2). Under the indirect dependency approach (Dayal 1993; Fujiwara 2021, a.o.), by contrast, the two *wh*-elements are not directly related syntactically. As shown in (7b), the embedded *wh*-phrase moves to the edge of the embedded CP but does not move further. The matrix clause contains its own independent *wh*-phrase, which moves to Spec-CP^{MAT} and is co-indexed with the embedded clause as a whole rather than with the embedded *wh*-phrase.

(7) a. Direct dependency approach



b. Indirect dependency approach



(Fujiwara 2021)

According to Fujiwara (2021), the two approaches make different predictions about the nature of the embedded clause. Under the direct

dependency approach, the embedded clause should be a declarative clause, since the *wh*-phrase moves out of it, leaving behind a trace in a configuration parallel to standard long-distance *wh*-movement. In Japanese, declarative embedded clauses are marked by the complementizer *to*. By contrast, under the indirect dependency approach, the embedded clause should retain its interrogative status, since the embedded *wh*-phrase stays within the embedded clause and is valued there by the local Q-particle *ka*.

The data support the indirect dependency approach. As shown in (6), the embedded clause in *wh*-scope-marking constructions is marked by the Q-particle *ka*, indicating that it is interrogative rather than declarative. Moreover, replacing *ka* with the declarative complementizer *to* yields an ungrammatical result, as shown in (8).

- (8) *Ano-hito-wa [dare-ga senkyo-ni tousensi-soo to]
 that-person-TOP who-NOM election-to win-EVID COMP
 nante itte-ta kke?
 what_{PRON} say-PST Q
 'What did that person say: who would win the election?'

The ungrammaticality of (8) demonstrates that the embedded clause in *wh*-scope-marking constructions must be interrogative. This is inconsistent with the direct dependency approach, which predicts a declarative embedded clause. The indirect dependency approach, on the other hand, correctly predicts this pattern: since the two *wh*-elements enter the derivation independently and the embedded *wh*-phrase is valued within the embedded clause, the embedded clause retains its interrogative status.

This background is relevant to the present study because one important learner pattern is the *wh-wh* type, in which *wh*-elements appear in both the matrix and embedded clauses, as in *What do you think who liked Mary?*. The surface form of this construction resembles a *wh*-scope-marking construction. As will be discussed in Section 2.4, previous analyses have treated the *wh-wh* type as involving a single derivational mechanism. The present study, however, argues that prosodic evidence can distinguish at least two structurally distinct productions that share this surface form.

2.3 Prosody as evidence for syntax

The present study adopts the view that prosody can provide evidence for syntactic structure. In particular, it focuses on the Intonational Phrase (IntP), which is taken here to be the prosodic domain most relevant to clause-level structure. Following Nespor and Vogel (1986), an IntP can be understood as the domain of an intonation contour, whose boundaries coincide with positions where pauses may be introduced in a sentence.

The present study builds on work on the syntax–prosody interface, especially Match Theory (Selkirk 2009, 2011) and Royer (2022)’s proposal that prosody can serve as syntactic evidence. Following the recursive organisation of the prosodic hierarchy, Match Theory proposes a mapping system from syntax to phonology, as shown in (9).

- (9) a. Match Clause
A clause in syntactic constituent structure must be matched by a constituent of a corresponding prosodic type in phonological representation, call it ι .
- b. Match Phrase
A phrase in syntactic constituent structure must be matched by a constituent of a corresponding prosodic type in phonological representation, call it Φ .
- c. Match Word
A word in syntactic constituent structure must be matched by a constituent of a corresponding prosodic type in phonological representation, call it ω .

Selkirk (2009:40)

In this framework, a clause corresponds to CP, a phrase corresponds to XP other than CP, and a word corresponds to X^0 in syntax. The original formulation of Match Theory permits a degree of mismatch between syntactic and prosodic structure, since syntactic constituents may fail to be matched by a corresponding prosodic constituent when higher-ranked markedness constraints are satisfied (Selkirk 2009). Under this

view, prosodic phrasing is not a direct and faithful reflection of syntax but may be modulated by other phonological factors.

The present study adopts a stronger view, developed by Royer (2022) under the label *Prosody as Syntactic Evidence*. Drawing on evidence from prosodic allomorphy in Mayan languages, Royer (2022) argues that apparent mismatches between syntax and prosody should not be attributed to phonological factors, but should instead prompt a revision of the syntactic analysis. On this approach, prosodic structure is taken to be a direct and transparent reflection of syntactic structure. The present study adopts this methodological stance and extends it to L2 production: where a prosodic boundary is observed, a syntactic boundary—specifically a CP boundary—can be inferred.

The correspondence assumed in the present study is schematised in (10). A single CP is mapped onto a single IntP, whereas two CPs are mapped onto two IntPs.

- (10) a. [CP ...] ↔ (t)
 b. [CP ...] [CP ...] ↔ (t) (t)

In the present study, IntP boundaries are identified on the basis of pause duration. Following Kandybowicz (2020), an IntP boundary is posited when the intervening pause exceeds 100ms.

This approach is applied to learner productions by treating IntP boundaries as evidence for clause-level syntactic structure. A production realised as a single IntP is consistent with a single-CP structure, whereas a production realised with two IntPs suggests a structure containing two CPs. Under this view, prosodic variation in learner speech is not merely a phonetic phenomenon but may reflect structural variation in L2 grammar: if prosody can reveal underlying syntactic structure, then learner productions that appear identical in surface form may nevertheless differ in their syntactic representations.

2.4 Previous studies on long-distance *wh*-questions by JLEs

Previous studies have shown that Japanese learners of English (JLEs) produce not only target-like long-distance *wh*-questions but also a range

of non-target-like forms. In this paper, target-like long-distance *wh*-questions are referred to as the *target type*. Productions in which *wh*-elements appear in both the matrix and embedded clauses, such as *What do you think who liked Mary?*, are referred to as the *wh-wh* type. Productions in which a *wh*-element appears only in the embedded clause, such as *Do you think who liked Mary?*, are referred to as the *do-wh* type. These categories are used descriptively in the discussion below.

Wakabayashi and Okawara (2003) investigated Japanese learners' production of long-distance *wh*-questions by means of an Elicited Production Task (EPT) adapted from Crain and Thornton (1998). In their study, participants were presented with contexts designed to elicit long-distance *wh*-questions. The target items included subject and object *wh*-questions with both *what* and *who*, such as *What do you think is in the bag?*, *Who do you think loved Mr. Yellow?*, *What do you think Mr. Yellow eats?*, and *Who do you think Mr. Yellow loved?*. The participants were 16 Japanese university students

The results revealed several production types. In addition to the target type, learners produced the *wh-do* type, the *wh-wh* type, and the *do-wh* type, as illustrated in (11). The *wh-do* type involved a matrix *wh*-phrase together with dummy *do* inside the embedded clause, the *wh-wh* type contained *wh*-elements in both the matrix and embedded clauses, and the *do-wh* type lacked a matrix *wh*-phrase altogether.

- (11) a. The target type
Who do you think loved Mr. Yellow?
- b. The *wh-do* type
Who do you think did Mr. Yellow kiss?
Intended: 'Who do you think Mr. Yellow kissed?'
- c. The *wh-wh* type
What do you think who loved Mr. Yellow?
Intended: 'Who do you think loved Mr. Yellow?'
- d. The *do-wh* type
Do you think who Mr. Yellow loved?

Intended: ‘Who do you think Mr. Yellow loved?’

To account for these non-target-like productions, Wakabayashi and Okawara (2003) argued that learner productions reflect two factors: differences in the strength of functional categories between Japanese and English, and the derivational economy principle Merge over Move (MoM) (Chomsky 1995). On their analysis, the *wh-wh* type arises because the learner merges a new *wh*-word in the matrix clause rather than moving the embedded *wh*-phrase all the way to the matrix clause. A schematic version of their analysis is given in (12).

(12) a. *wh-wh* type

What do you think who loved Mr. Yellow?

b. Analysis proposed by Wakabayashi and Okawara (2003)

[_{CP} What [_C do] [_{TP} you think [_{CP} who_i [_{TP} ~~who~~_i loved Mr. Yellow]]]]



The *do-wh* type is analyzed differently. In this case, the main clause C contains a strong [Q] feature, which triggers movement of dummy *do* to the matrix clause. However, because the [+*wh*] feature is not strong enough to attract the embedded *wh*-phrase, the *wh*-phrase remains in the embedded clause. The result is a sentence in which interrogative force is marked in the matrix clause but the *wh*-word appears only inside the embedded clause. The important point for the present study is that Wakabayashi and Okawara (2003) treated learner productions as syntactically systematic and attempted to derive them within a generative framework. At the same time, their analysis was based primarily on surface word order.

Kimura (2022) also investigated the acquisition of long-distance *wh*-questions by Japanese learners of English (JLEs), with a particular focus on how the internal configuration of *wh*-expressions differs between learners and native speakers. Here we focus only on the JLE results. Kimura used both an Acceptability Judgment Task and an Elicited Production Task. In the production task, each item consisted of a short contextual sequence followed by an instruction requiring participants to form a question using

a given *wh*-word, a main verb, and an embedded verb. A representative example is given in (13).

- (13) a. **Slide 1:** You heard that Ms. Red loves someone.
- b. **Slide 2:** However, Mr. Green doesn't know who it is.
- c. **Slide 3:** So, please ask him a question about the person that Ms. Red loves.
{who, love, think}
- d. **Target:** Who do you think (that) Ms. Red loves?

A notable property of Kimura's materials is that only *who* was used as the *wh*-word. This design avoided the possibility that learners would rely on formulaic expressions such as *What do you think...?*, and it also avoided ambiguity between true scope-marking uses of *what* and possible spell-out of intermediate copies.

The results are summarized in Table 1. Native speakers of English (NSEs) produced no ungrammatical sentences. In contrast, Japanese learners of English (JLEs) produced non-target-like forms. Lower-intermediate JLEs (JLE-L) produced both the *wh-wh* type and the *do-wh* type, whereas upper-intermediate JLEs (JLE-U) produced only the *wh-wh* type.

Table 1: EPT results for JLEs (Kimura 2022:233, Table 6-12, slightly modified)

	NSEs (<i>n</i> = 147)	JLE-L (<i>n</i> = 85)	JLE-U (<i>n</i> = 64)
Target	140 (95%)	37 (44%)	55 (86%)
<i>wh-wh</i>	0 (0%)	20 (24%)	8 (13%)
<i>do-wh</i>	0 (0%)	20 (24%)	0 (0%)
<i>wh-in-situ</i>	0 (0%)	0 (0%)	0 (0%)
<i>wh-copying</i>	0 (0%)	0 (0%)	0 (0%)

Based on these results, Kimura (2022) argued that JLEs and NSEs differ in the internal configuration of *wh*-expressions. Adopting Cable’s (2010) analysis, Kimura represented this difference as in (14). In the JLE interlanguage, shown in (14a), the Q feature adjoins to NP, yielding an NP projection. By contrast, in the native grammar, shown in (14b), the Q feature merges with NP and projects a QP.

- (14) a. $[_{NP} [_{NP} \text{ who}] [_Q \text{ } uOp: ___]]$
b. $[_{QP} [_{NP} \text{ who}] [_Q \text{ } uOp: ___]]$

The configuration (14a) allows the NP and Q to move independently, as illustrated in (15). The Q feature moves to the matrix C position for valuation, establishing interrogative force. The lexical *wh*-word also moves, but unlike English *wh*-movement, this movement is treated as scrambling, which is available in the learner’s L1 grammar.

- (15) $[_{NP} \text{ who}] [_C \text{ } iOp] [...[_{NP} [_{NP} \text{ who}] [_Q:uOp]]]]$

This analysis allows Kimura to derive both the *wh-wh* type and the *do-wh* type. In the *wh-wh* type, the lexical *wh*-word moves to the embedded CP via scrambling, while the [*uOp*] feature moves to the matrix CP and

is spelled out as *what*, functioning as a scope marker, as shown in (16) (Kimura 2022). In the *do-wh* type, the same two movements apply, but the operator feature is not spelled out as a matrix *wh*-word, yielding the pattern in (17).

- (16) $[_{CP}^{Mat} \text{ What}_{[uOp]} \text{ do you think } [_{CP}^{EMB} \text{ who Mr.Blue kicked } t]]?$
(Kimura 2022:247)

- (17) $[_{CP}^{Mat}_{[Op]} \text{ Do you think } [_{CP}^{EMB} \text{ who Mr.Blue kicked } t]]?$
(Kimura 2022:246)

Kimura further argued that target-like productions can also be derived under this analysis, as shown in (18). Crucially, this does not mean that learners have acquired native-like long-distance *wh*-movement. Rather, a target-like surface order may be generated through matrix scrambling of the lexical *wh*-phrase together with movement of the operator feature. In this sense, even apparently target-like productions may reflect a learner-specific derivation.

- (18) $[_{CP}^{MAT} \text{ Who}_{[Op]} \text{ do you think } [_{CP}^{EMB} \text{ Mr. Blue kicked } t]]?$

Kimura's analysis is thus fundamentally different from that of Wakabayashi and Okawara (2003). Whereas Wakabayashi and Okawara (2003) explained learner productions in terms of feature strength and the economy principle Merge over Move, Kimura (2022) attributed them to differences in the internal structure of *wh*-expressions, which allows the lexical *wh*-word and the operator feature to behave independently.

Other studies have also addressed the *wh-wh* type from different theoretical perspectives. Yamane (2003) proposed a parameter-based account in which JLEs adopt a German-type grammar that permits partial *wh*-movement, leading to the production of forms with *wh*-elements in both clauses. Schulz (2011), on the other hand, argued that at least some *wh-wh* productions reflect processing difficulty rather than a syntactic deficiency. On this view, the *wh-wh* type may arise when learners fail

to integrate the embedded *wh*-phrase into a single-clause structure under processing pressure. Together, these studies demonstrate that the *wh-wh* type has attracted multiple competing explanations within the GenSLA literature.

At the same time, all of these studies share an important limitation for the purposes of the present study. Although they provide valuable syntactic accounts of learner productions, they focus primarily on surface word order. They do not address the possibility that productions with the same surface form may differ in underlying syntactic structure. The present study builds on this line of research and offers a new perspective: prosodic structure can reveal underlying syntactic variation even within superficially similar learner productions.

3. RESEARCH QUESTIONS AND PREDICTIONS

Previous studies have shown that Japanese learners of English (JLEs) produce not only target-like long-distance *wh*-questions but also non-target-like forms such as the *wh-wh* type and the *do-wh* type. These studies have offered important syntactic accounts of such productions within the framework of Generative Second Language Acquisition. However, two issues remain to be explored further.

First, previous studies have focused on surface form and have paid no attention to prosody. If prosody can serve as evidence for underlying syntactic structure, as proposed by Royer (2022), then prosodic analysis may reveal structural differences even among learner productions that appear identical on the surface. Second, learner productions should be examined not only at the group level but also at the individual level. Although learners with the same L1 are often discussed as if they formed a uniform group, interlanguage grammars may vary considerably across individuals (Wakabayashi and Takahashi 2025). For this reason, the present study pays close attention to individual learner patterns as well as overall tendencies.

3.1 Research Questions

Based on these considerations, this study addresses the following research questions:

- (19) a. **RQ1:** What kinds of long-distance *wh*-question productions do JLEs produce?
- b. **RQ2:** Are target-like productions prosodically uniform?
- c. **RQ3:** What does prosodic variation reveal about the underlying syntactic structure of learner productions?

RQ1 asks about the range of production types in the learner data, replicating the descriptive focus of previous studies. RQ2 directly addresses the central claim of the present study: if prosody reflects underlying syntax, then target-like surface forms need not be prosodically uniform. RQ3 asks what structural conclusions can be drawn from observed prosodic variation, connecting the empirical findings to the theoretical framework introduced in Section 2.

3.2 Predictions

The present study adopts Match Theory (Selkirk 2011) and the stronger view that prosody can serve as syntactic evidence (Royer 2022). Under this approach, prosodic structure is assumed to reflect syntactic structure in a transparent way. More specifically, a single CP is expected to correspond to a single IntP, whereas two CPs are expected to correspond to two IntPs, as schematized in (10).

On this basis, two prosodic patterns are predicted for target-like long-distance *wh*-questions. Consider a participant who constructs a two-CP structure for a target-like production, with an empty category in the embedded clause. In this structure, embedded CP merges with main clause as adjunct clause. Under the prosody–syntax correspondence, this structure is predicted to be mapped onto two IntPs, as in (20a). Alternatively, consider a participant who constructs a single-CP structure, in which the *wh*-phrase moves to the matrix CP through embedded CP. In this structure, the embedded clause merges with matrix verb as its complement.

This structure is predicted to be mapped onto a single IntP, as in (20b). Both patterns are therefore expected in the learner data*.

- (20)
- a. $[_{CP}[_{CP} \text{ Who do you think }][_{CP} ec \text{ likes Mary}]]]$?
 - b. $(\text{Who do you think})_t (\text{likes Mary})_t$
 - c. $[_{CP} \text{ Who do you think } [_{CP} ec \text{ likes Mary}]]]$?
 - d. $(\text{Who do you think likes Mary})_t$

A further prediction concerns the *wh-wh* type. If this type involves two CPs—one for the matrix clause and one for the embedded clause—the utterance is predicted to be realized with two IntPs, as in (21a). By contrast, a one-IntP realization of the *wh-wh* type is not predicted under the initial hypothesis. This is because a *wh-wh* production realized as a single IntP would suggest a single-CP structure. However, within a single CP, the presence of two *wh*-phrases is not straightforwardly derivable, since each *wh*-phrase is normally associated with a CP position. The one-IntP *wh-wh* pattern is therefore unexpected, as shown in (21b).

- (21)
- a. $[_{CP}[_{CP} \text{ What do you think }][_{CP} \text{ who likes Mary}]]]$?
 - b. $(\text{What do you think})_t (\text{who likes Mary})_t$
 - c. $[_{CP} \text{ What do you think who likes Mary}]]]$?
 - d. $(\text{What do you think who likes Mary})_t$

These predictions are summarized in Table 2. The only pattern that is not predicted under the initial hypothesis is a *wh-wh* production realized as a single IntP. If such a pattern is nonetheless observed in the learner data, it will require an analysis that goes beyond the initial predictions. As shown in Section 6, this is indeed what the data reveal, providing further evidence for the claim that prosody can uncover structural variation that surface form alone cannot detect.

* Thanks to an anonymous reviewer for clarifying this point.

Table 2: Predicted prosodic patterns

	[CP Wh...]	[CP Wh...] [CP ...]
One IntP	(Wh...) ι	*(Wh...wh...) ι
Two IntPs	(Wh...) ι (...) ι	(Wh...)(wh...) ι

4. METHOD

4.1 Design

The present study employed an Elicited Production Task (EPT). The aim was to elicit participants' production of long-distance *wh*-questions and to examine their prosodic patterns. Because long-distance *wh*-questions rarely occur in spontaneous speech or ordinary conversation, the EPT was designed to elicit these structures in a controlled manner. The task was implemented using Gorilla Experiment Builder (Anwyl-Irvine et al. 2020), which enabled online data collection and allowed participants to complete the experiment on their own computers. Audio responses were recorded directly through the Gorilla platform.

4.2 Materials

The EPT consisted of 12 target contexts and 11 filler contexts. Each context was presented through a series of slides. In (22a), the event and its characters were introduced. In (22b), one character's perspective was presented. In (22c), another character expressed a contrasting view. In (22d), participants were required to formulate a question directed at the second character, using pre-selected words from the model answer. The presentation order of these prompt words was randomized.

- (22) a. Today, someone fought with Tom because the person hates him.
You, Mary, and Kate are talking about it.
- b. Mary thinks Tom's teacher hates him.
- c. But Kate has a different idea.

d. So please ask Kate about her idea. *Tom, think, hate*

Model Answer: Who do you think hates Tom?

The target items used either *think* or *believe* as the main verb. The target sentences with these verbs require *wh*-movement from their complement clauses to the sentence initial position, and do not allow the *wh*-phrase to remain *in situ*. Each target item varied in extraction position: either from the subject or object position of the embedded clause. Three items were created for each of the four conditions, yielding a total of 12 target items. The conditions and their expected responses are summarized in Table 3.

Table 3: Target items and expected responses

	Subject extraction	Object extraction
think	<i>Who do you think likes Mary?</i>	<i>Who do you think Mary likes?</i>
believe	<i>Who do you believe likes Mary?</i>	<i>Who do you believe Mary likes?</i>

The 11 filler items were included to prevent participants from developing a strategy specific to long-distance *wh*-questions. They consisted of a range of question types that did not require long-distance *wh*-movement.

4.3 Participants

The participants consisted of 11 Japanese learners of English (JLEs) and 11 native speakers of English (NSEs). The JLEs were undergraduate students or office workers at the time of data collection, with English proficiency levels ranging from elementary to upper intermediate as measured by the Quick Placement Test (QPT). The QPT scores and proficiency levels of the JLE participants are summarized in Table 4.

Table 4: QPT scores and proficiency levels of the JLE participants

Participant	QPT score	Level
jle005	46 / 64	Upper intermediate
jle007	46 / 64	Upper intermediate
jle006	39 / 64	Lower intermediate
jle011	38 / 64	Lower intermediate
jle001	38 / 64	Lower intermediate
jle003	38 / 64	Lower intermediate
jle008	33 / 64	Lower intermediate
jle010	31 / 64	Lower intermediate
jle009	30 / 64	Lower intermediate
jle002	28 / 64	Elementary
jle004	24 / 64	Elementary

The NSE data were collected through the Prolific platform. All 11 NSE participants were from the United States and were monolingual English speakers.

4.4 Procedure

The JLE participants first completed the QPT, which took approximately 30 minutes. Following the QPT, they completed the main task, which consisted of the target and filler items and took approximately 20 minutes. The NSE participants completed only the main task. All participants completed the experiment online using Gorilla Experiment Builder at home on their personal computers, and their spoken responses were recorded directly through the platform.

4.5 Prosodic Analysis

The prosodic analysis focused on whether the boundary between the matrix clause and the embedded clause was realized as a single Intonational Phrase (IntP) or as two separate IntPs. Audio recordings were examined using Praat (Boersma and Weenink 2024). Following the assumption

adopted in this study that prosodic structure can serve as evidence for underlying syntactic structure, IntP boundaries were used as a diagnostic for clause-level syntactic structure.

We used pause duration as a diagnosis to detect IntPs. When the pause between the matrix clause and the embedded clause was 100 ms or longer, the two clauses were analyzed as belonging to separate IntPs. This threshold follows Kandybowicz (2020), based on Crystal and House (1988).

5. RESULTS

5.1 Native speaker baseline

Table 5 shows the numbers of target and other responses produced by the NSEs. Although some participants produced non-target responses, such as yes/no questions (e.g., *Do you think Kate likes Jack?*) and declarative sentences (e.g., *I think Kate likes Jack.*), none of their utterances contained the *wh-wh* type or other non-target-like forms involving *wh*-elements in the embedded clause.

	Target	Others (e.g. yes/no questions, declaratives, unanalyzable responses)
nse_001	12	0
nse_002	12	0
nse_006	12	2
nse_011	12	0
nse_005	11	1
nse_003	10	2
nse_009	9	3
nse_008	7	5
nse_010	7	5
nse_007	3	9
nse_004	1	11

With respect to prosody, all target-like productions by the NSEs were realized as a single IntP. No pause exceeding 100 ms was observed at the matrix-embedded clause boundary, and no pitch reset was detected in any of the NSE target-like productions. An example is shown in Figure 1. In the production of *Who do you believe Kate criticized?*, the F0 contour shows a continuous decline across the matrix-embedded clause boundary, with no pause or pitch reset. This prosodic pattern is consistent with a single-CP structure, as predicted under the prosody-syntax correspondence adopted in this study.

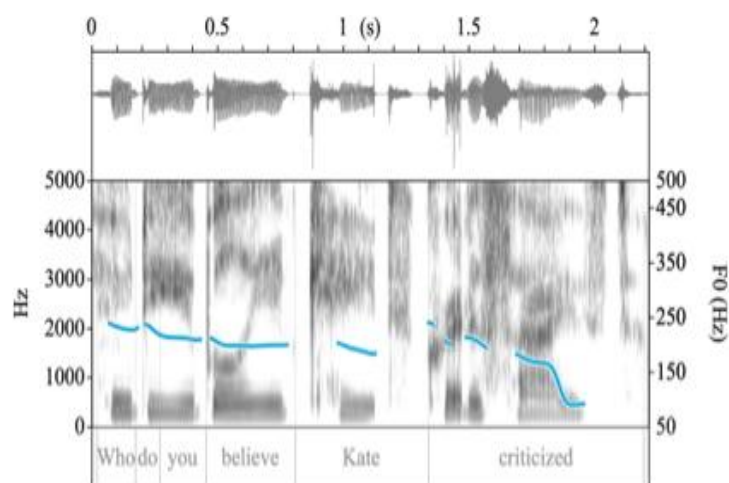


Figure 1: Production of *Who do you believe Kate criticized?* by a NSE[†]

The NSE baseline thus establishes that native speakers consistently realize long-distance *wh*-questions as a single IntP, in line with the prediction that a single-CP structure maps onto a single IntP.

5.2 Learner productions by surface form

[†] An anonymous reviewer pointed out the F0 rise at the onset of the word *criticized* indicates the onset of another IntP.

Table 6 presents the surface forms produced by the JLEs. The responses are categorized by production type, following the classification introduced in Section 2.4. The what-who type and the how-who type are grouped together as the *wh-wh* type, while the do-who type is treated as the *do-wh* type.

Table 6: Surface forms produced by the JLEs

Participants	QPT score	Level	Subject				Object			
			Target	who-do	what-who	how-who	do-who	Target	who-do	what-who
jle005	46 / 64	Upper intermediate	4					6		
jle007	46 / 64	Upper intermediate	5					2		
jle006	39 / 64	Lower intermediate	6						5	
jle011	38 / 64	Lower intermediate	6					4		
jle001	38 / 64	Lower intermediate	5					4		
jle003	38 / 64	Lower intermediate						9		
jle008	33 / 64	Lower intermediate			2			1		1
jle010	31 / 64	Lower intermediate			1	3				
jle009	30 / 64	Lower intermediate	4					1		
jle002	28 / 64	Elementary			4		2			
jle004	24 / 64	Elementary			5		2			2

Several patterns are worth noting. First, with the exception of jle008, participants who produced non-target forms did not produce any target-like utterances in the same extraction condition. For example, jle003 produced only the *do-wh* type, and jle010 produced only the *wh-wh* type. Second, the two elementary learners, jle002 and jle004, each produced both the *wh-wh* type and the *do-wh* type, whereas the lower-intermediate learners who produced non-target forms produced either the *wh-wh* type or the *do-wh* type, but not both. Third, the upper-intermediate learners, jle005 and jle007, produced only target-like sentences, whereas lower-intermediate learners showed more variation, with some producing target-like sentences exclusively and others producing only non-target forms. These patterns suggest a loose correlation between proficiency level and production type, though individual variation is also evident within the lower-intermediate group.

5.3 Prosodic variation in learner productions

We next examine prosodic variation in the learner data. We begin with the target-like productions. Figure 2 shows the production of *Who do you believe Tom respects?* by jle007. The F0 contour shows a continuous pattern across the matrix–embedded clause boundary, with no pause or pitch reset. This utterance was therefore analyzed as a single IntP, consistent with the NSE baseline.

However, not all target-like productions showed this pattern. Figure 3 shows the production of *Who do you believe likes Jack?* by jle001. In this case, a pause is observable between the matrix clause *Who do you believe* and the embedded clause *likes Jack*, and the F0 level at the onset of the embedded clause is higher than at the end of the matrix clause, indicating pitch reset. This utterance was therefore analyzed as consisting of two IntPs, despite being target-like in surface form.

Table 7 summarizes the target-like productions by IntP pattern across the JLE group. The data reveal considerable individual variation. A comparison between jle011 and jle005 is particularly informative: although both participants produced ten target-like utterances, jle011 produced both one-IntP and two-IntP utterances, whereas jle005 produced exclusively two-IntP utterances. This contrast shows that target-like surface form does not correspond to a uniform prosodic pattern, even within the same proficiency level.

Table 7: Target-like productions and IntP patterns in the JLE group

	<i>(wh...)I</i>	<i>(wh...)I (ec...)I</i>	Sum
jle011	4	6	10
jle005	0	10	10
jle001	4	5	9
jle007	5	1	6
jle006	4	2	6
jle009	2	3	5

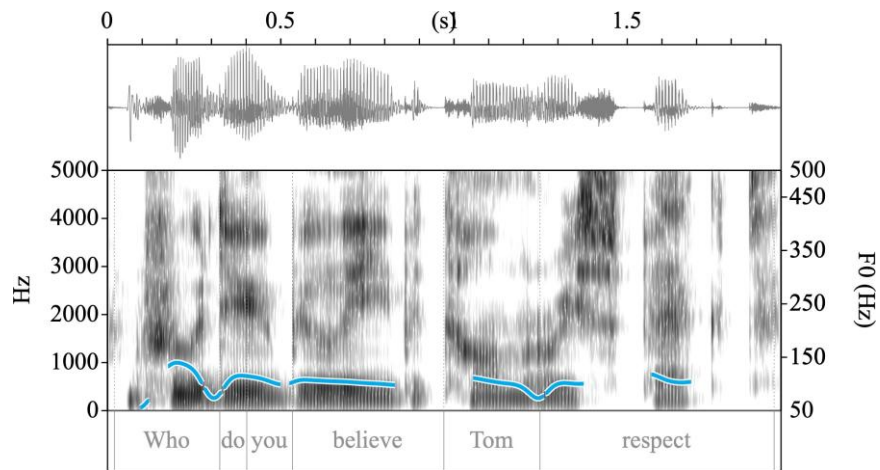


Figure 2: Production of *Who do you believe Tom respects?* by jle007

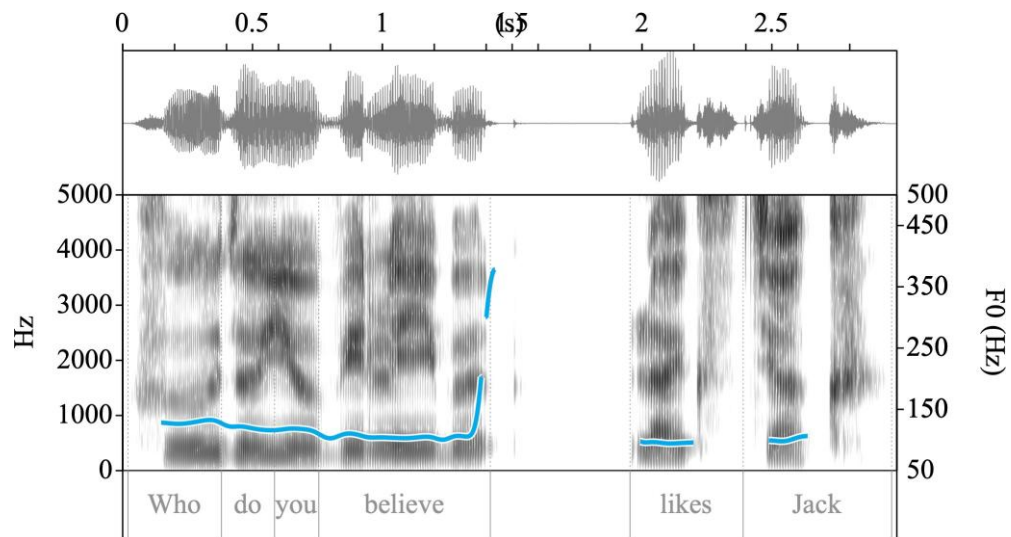


Figure 3: Production of *Who do you believe likes Jack?* by jle001

We now turn to the *wh-wh* type. Figures 4 and 5 show productions in which a pause is observable between the matrix and embedded clauses, accompanied by pitch reset. These utterances were therefore analyzed as consisting of two IntPs.

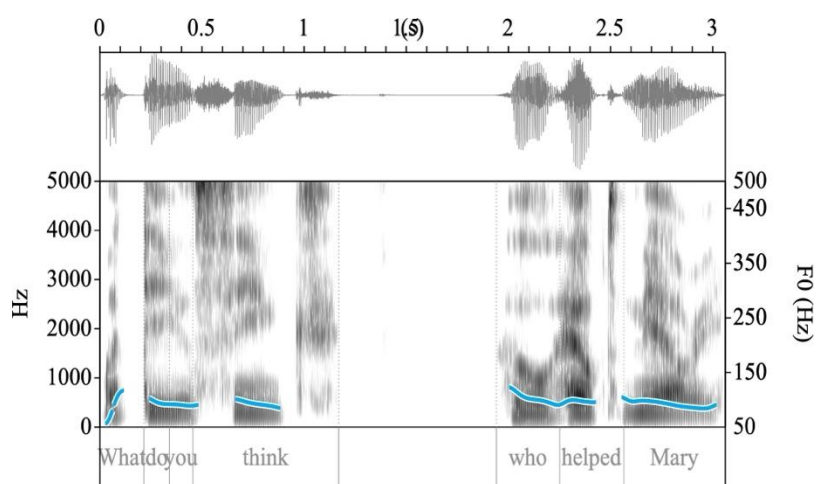


Figure 4: Production of *What do you think who helped Mary?* by jle010

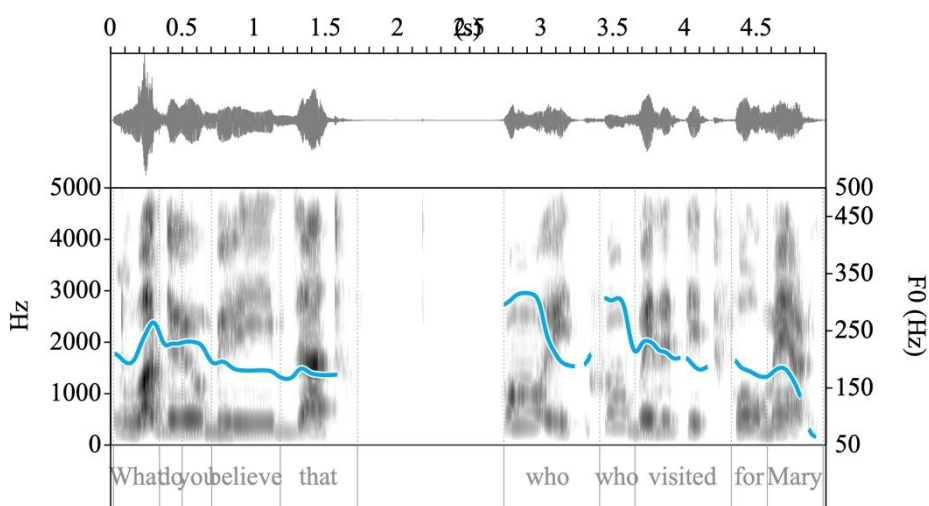


Figure 5: Production of *What do you believe that who visited Mary?* by jle004

A different prosodic pattern was also observed. Figure 6 shows a *wh-wh* production by jle010 in which no pause or pitch reset appears at the matrix–embedded clause boundary. The entire utterance was therefore analyzed as a single IntP. This pattern is unexpected under the initial prediction, since two *wh*-elements are present within a single IntP.

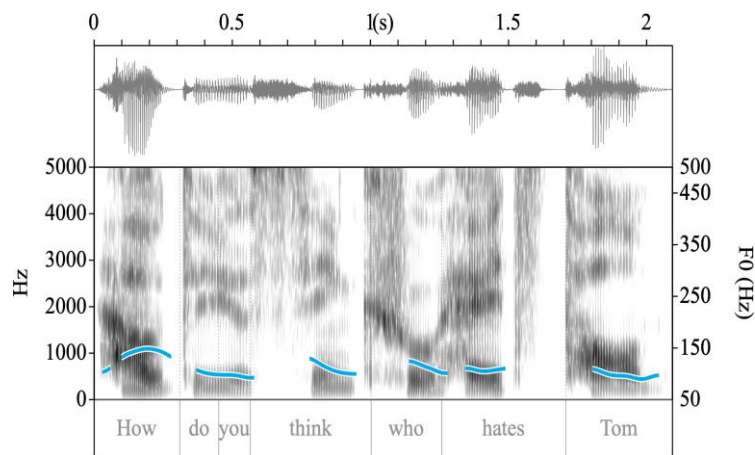


Figure 6: Production of *How do you think who hates Tom?* by jle010

A further *wh-wh* pattern was observed in the data. Figure 7 shows a production by jle002 in which a pause appears between *who* and *hit*, with the second *wh*-word grouped prosodically with the matrix clause rather than the embedded clause. This utterance was analyzed as consisting of two IntPs, with the boundary falling after the second *wh*-element.

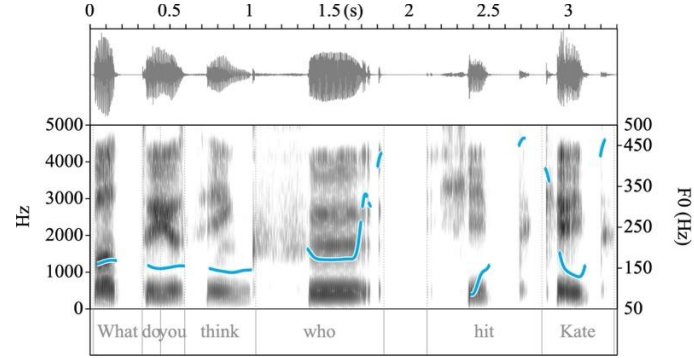


Figure 7: Production of *What do you think who hit Kate?* by jle002

Table 8 summarizes the *wh-wh* productions and their IntP patterns. Whereas jle004 produced only two-IntP utterances, jle002 and jle010 produced both one-IntP and two-IntP utterances, indicating that the *wh-wh* type is not prosodically uniform either.

Table 8: *wh-wh* productions and IntP patterns in the JLE group

	$(wh..wh.)_I$	$(wh...)_I (wh..)_I$	Sum
jle004	0	6	6
jle002	1	3	4
jle010	2	2	4

Table 9 provides a summary of the prosodic patterns observed in the learner data. Four of the five patterns are consistent with the predictions made in Section 3. The only unexpected pattern is the one-IntP realization of the *wh-wh* type, which was nonetheless attested in the data. The implications of this finding are taken up in Section 6.

Table 9: Summary of the learner results

Group	Pattern	Position of <i>wh</i>	Number of IntPs	Prediction	Result
NS	Target	<i>Wh...ec...</i>	1	✓	✓
JLE	Target-like (1)	<i>Wh...ec...</i>	1	✓	✓
JLE	Target-like (2)	<i>Wh...ec...</i>	2	✓	✓
JLE	Non-target-like	<i>Wh...wh...</i>	2	✓	✓
JLE	Non-target-like	<i>Wh...wh...</i>	1	×	✓

6. DISCUSSION

6.1 What prosody reveals about learner syntax

Let us first clarify how prosodic patterns are interpreted in the present study. Some may wonder if prosodic data can really provide us with insights into learners' syntactic knowledge. One objection is that pauses in learner production merely reflect disfluency rather than syntactic structure. Under this view, a learner who pauses between the matrix and embedded clauses is simply hesitating, not producing a structurally motivated boundary. However, the pause patterns observed in the present data cannot be reduced to disfluency alone. If pauses were simply a by-product of disfluent production, they would be expected to occur randomly throughout the utterance—within clauses as well as between them. In the present data, by contrast, pauses appear systematically at the boundary between the matrix clause and the embedded clause, not at other positions within the utterance. This distributional pattern is inconsistent with a disfluency account and instead suggests that the pauses reflect prosodic structuring at clause boundaries.

On this basis, the present study treats prosodic patterns as evidence for underlying syntactic structure. More specifically, a single IntP is taken to correspond to a single CP, whereas two IntPs are taken to correspond to two CPs, following the prosody–syntax correspondence schematised in Section 2. Under this view, prosodic variation in learner speech is not merely phonetic variation but structural variation in learner grammar. A learner who consistently produces a given construction as a single IntP is treated as operating with a single-CP structure, whereas a learner who

produces the same construction with two IntPs is treated as operating with a two-CP structure.

This perspective has a direct consequence for the analysis of learner productions. A sentence that appears target-like on the surface does not necessarily reflect a target-like syntactic representation. Prosody provides an additional dimension of evidence that surface form alone cannot supply. As shown below, this consequence bears on the interpretation of both target-like productions and non-target-like productions in the present data.

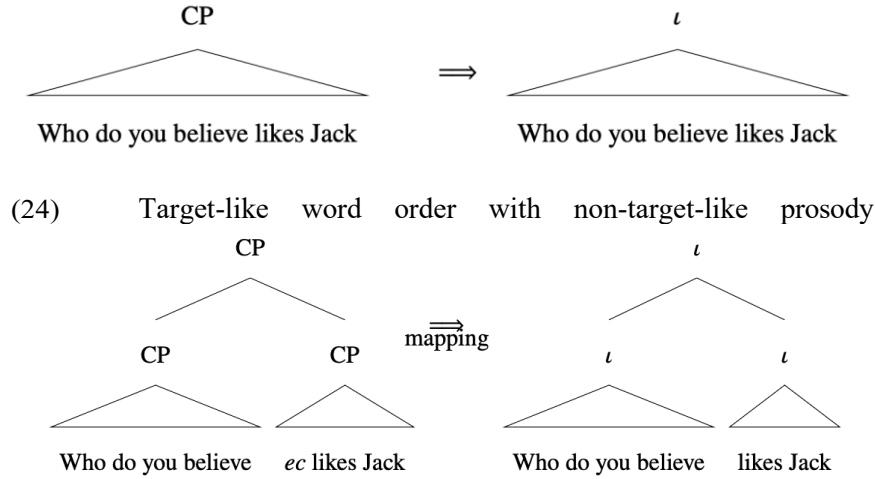
6.2 Why target-like form does not always imply target-like syntax

A central finding of the present study is that target-like surface form does not always imply target-like syntax. This finding challenges a common assumption in GenSLA research. In many previous studies, when learners produced a target-like sentence, the underlying representation was assumed to be target-like as well (Hawkins and Hattori 2006). On that view, a sentence such as *Who do you believe likes Jack?* would be analyzed as involving a single CP derived by obligatory and successive-cyclic *wh*-movement, just like the corresponding native-speaker production. The present data suggest that this assumption is not warranted.

As shown in Table 7, productions with the target-like word orders utterances in the JLE group were not prosodically uniform. Some were produced as a single IntP, whereas others were produced as two IntPs. Under the prosody–syntax correspondence adopted here, these two prosodic patterns suggest different underlying structures.

This point is illustrated schematically in (23) and (24). In (23), the target-like sentence is produced with a single IntP, which corresponds to a single CP. In (24), by contrast, the same surface string is produced with two IntPs. On the present analysis, this prosodic pattern suggests that the learner constructs two CPs, despite producing a target-like sentence on the surface.

- (23) Target-like word order with target-like prosody



The contrast between these two patterns is also reflected in the differences among individual learners. As shown in Table 10, jle005 produced ten target-like utterances, all of which were realized with two IntPs. This is a striking result: not a single production by jle005 showed the single-IntP pattern found in native speaker data. By contrast, jle007 produced five one-IntP and one two-IntP utterance, a pattern much closer to the native speaker baseline. These individual differences suggest that learners differ not only in whether they produce target-like surface forms but also in the underlying syntactic structures they employ to generate those forms.

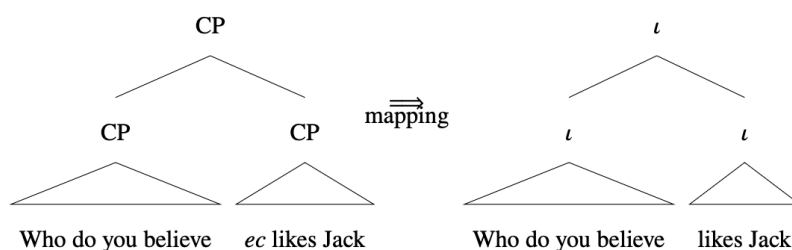
Table 10: Target-like productions and IntP patterns in the JLE group (repeated)

	$(wh...)I$	$(wh...)I (ec...)I$	Sum
jle011	4	6	10
jle005	0	10	10
jle001	4	5	9
jle007	5	1	6
jle006	4	2	6
jle009	2	3	5

6.2.1 Apparently target-like productions with two CPs

The two-IntP target-like productions require a structural analysis that goes beyond the simple assumption of a single CP. Under the prosody–syntax correspondence adopted here, a two-IntP realization implies a structure containing two CPs. One possible derivation involves an empty category in the embedded clause, as schematised in (25). In this structure, the matrix clause and the embedded clause each project a CP, and the two CPs are mapped onto two IntPs respectively.

(25) Apparently target-like production with two CPs



The present experiment does not allow us to determine the precise status of the empty category. At least two possibilities remain open, as shown in (26). The empty category may be a null pronoun *pro* coindexed with the matrix *wh*-word, as in (26a), or a deleted copy of the *wh*-word as in (26b).

- (26) a. [_{CP}Who_i do you believe] [_{CP}*pro*_i likes Jack?]
 b. [_{CP}Who_i do you believe] [_{CP}~~who~~_i likes Jack?]

If the empty category is a deleted copy as in (26b), the derivation might seem identical to the target-like derivation described in Section 2, in which the *wh*-phrase moves successively through the embedded Spec-CP to the matrix Spec-CP. However, this is not the case. Under the present analysis, the matrix CP and the embedded CP are independent projections: the verb *think* does not take the embedded CP as its argument. Successive-cyclic movement presupposes a syntactic dependency between the two clauses,

but this dependency does not hold in a structure where the two CPs are independent. This rules out the deleted copy analysis.

The empty category must therefore be *pro*, as in (26a). On this analysis, the learner constructs an independent embedded CP containing a null pronoun co-indexed with the matrix *wh*-word as a bound variable, in the same way as null pronouns operate in Japanese. Since *pro* is base-generated in the embedded clause rather than derived through movement, no successive-cyclic movement is required. The surface form is target-like, but the underlying structure is genuinely non-target-like.

The crucial point for the present argument is that the two-IntP prosodic pattern reveals this non-target-like structure, which surface word order alone cannot detect. Two IntPs indicate two CPs, and two CPs indicate that the embedded clause has been projected as a full CP containing *pro*. This is information that surface word order alone cannot provide, and it shows that the learner's grammar is more complex than a simple surface-form analysis would suggest.

A further question is why some learners produce target-like utterances with one IntP while others produce them with two IntPs. The optionality observed in the data—particularly in the case of jle011, who produced both patterns—suggests that the presence or absence of the embedded CP projection may vary even within a single learner's grammar. Following Wakabayashi (1997, 2009), this variability can be understood in terms of optionality in numeration: the learner may optionally select an empty category for the embedded clause. When this option is selected, the derivation yields a two-CP structure mapped onto two IntPs. When it is not selected, the derivation yields a single-CP structure mapped onto a single IntP. The co-existence of both patterns in a single learner's data is therefore expected under this account.

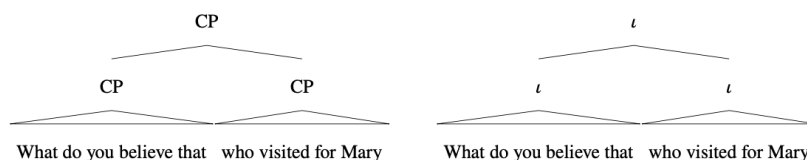
6.2.2 The *wh-wh* type

The *wh-wh* type also provides evidence that learner productions are not uniform at the level of underlying syntax. As shown in the previous section, the *wh-wh* productions in the learner data were realized with more than one prosodic pattern. Some were produced with two IntPs, whereas others were produced with a single IntP. This variation suggests that the *wh*-

wh type cannot be treated as a single construction with a uniform underlying derivation.

We begin with the *wh-wh* productions realized with two IntPs. Under the prosody–syntax correspondence adopted in this study, this prosodic pattern suggests a structure containing two CPs. This analysis is illustrated in (27). In this structure, the sentence-initial *wh*-phrase is merged in the matrix clause, while the second *wh*-phrase is merged in the embedded clause. The resulting structure is consistent with the indirect dependency approach to *wh*-scope-marking constructions introduced in Section 2.

(27) The *wh-wh* type with two IntPs



This analysis is compatible with previous claims that the *wh-wh* type in L2 English reflects Merge over Move (Wakabayashi and Okawara 2003). On that view, the sentence-initial *wh*-phrase is merged directly in the matrix clause rather than derived by movement from the embedded clause. The present data are consistent with this account: the two-IntP pattern confirms that the matrix and embedded clauses constitute separate CPs, and the matrix *wh*-phrase is not a moved copy of the embedded *wh*-phrase but an independently merged element. At the same time, the present data suggest that Merge over Move alone may not be sufficient to explain the full range of learner patterns.

A further factor appears to be L1 transfer. As discussed in Section 2, Japanese allows *wh*-scope-marking constructions in which a matrix *wh*-phrase functions as a scope marker for an embedded *wh*-phrase. This makes it plausible that Japanese learners draw on this L1 option when producing the *wh-wh* type in English. Crucially, the L1 evidence supports this interpretation. As shown in (6), Japanese *wh*-scope-marking constructions permit more than one type of matrix scope marker: *nante*, which corresponds semantically to *what*, and *doo*, which corresponds to *how*. If L1 transfer plays a role in learner productions, both *what*-type and

how-type matrix *wh*-phrases would be expected to appear in sentences like (28).

- (28) a. What do you think who helped Mary?
 b. How do you think who hates Tom?

This prediction is borne out in (28). The occurrence of *how* as a scope-marking element in learner English is particularly telling. It is difficult to attribute this pattern to Merge over Move alone, since there is no obvious reason why English would license *how* in a scope-marking function. The most natural explanation is that the learner is drawing on the Japanese option in which *doo* (corresponding to *how*) functions as a scope marker. The learner data are therefore more naturally explained if both Merge over Move and L1 transfer are taken into account. This finding also stands in contrast to Schulz (2011), who argued against an L1 transfer account of the *wh-wh* type. The present data suggest that L1 transfer must be taken seriously as a contributing factor.

At the same time, not all *wh-wh* productions fit this two-CP pattern. As shown in the previous section, some *wh-wh* utterances were produced as a single IntP. These cases cannot be analyzed in the same way as the two-IntP pattern and require a different account. I turn to this issue in the next section.

6.2.3 Productions without an embedded CP layer

Some learner productions cannot be analyzed as involving two CPs. In particular, a subset of the *wh-wh* type was produced as a single IntP. This pattern is unexpected under the initial prediction, because two *wh*-elements would normally suggest two CP-related positions. The one-IntP realization therefore forces a different structural analysis. Under the prosody–syntax correspondence adopted in this study, a single IntP maps to a single CP. This pattern therefore suggests that the embedded clause does not project a CP layer. I propose that these productions involve an embedded TP rather than an embedded CP (Bošković 1997).

This pattern is illustrated by the learner productions discussed in Section 5. In these examples, no pause appears between the matrix clause

and the embedded clause, indicating that the entire utterance was produced as a single IntP. If one IntP corresponds to one CP, these productions cannot be analyzed in the same way as the two-IntP *wh-wh* type. A different derivation is therefore required. The proposed structure is shown in (29). On this analysis, the embedded clause is a TP, and no embedded CP is projected. As a result, the embedded *wh*-phrase cannot be valued by a local C head within the embedded clause.

(29) The *wh-wh* type without an embedded CP layer



A possible derivation is given in (30). In this structure, the embedded clause is projected only as TP. Because no embedded C head is present, no valuation of the embedded *wh*-phrase occurs within the embedded clause. As a result, the [*uOp*] feature of the embedded *wh*-phrase is free to move to the matrix CP, where it is valued and spelled out as the matrix *wh*-word (Kimura 2022). This derivation follows the direct dependency approach to *wh*-scope-marking constructions, in which the feature of the embedded *wh*-phrase moves directly to the matrix clause.

(30) [_{CP}How[*uOp*] do you think [_{TP}who[*uOp*] hates Tom]]

One might ask why the direct dependency approach is adopted here, given that the indirect dependency approach was supported for *wh*-scope-marking constructions in Section 2. The answer lies in the structural difference between the two cases. As argued in Section 2, the indirect dependency approach is supported by the fact that the embedded *wh*-phrase is valued by the local Q-particle *ka* in the embedded clause, which creates a freezing effect that prevents further movement. In the present case, however, the embedded clause lacks a CP layer entirely. Without an

Under this view, the one-IntP *wh-wh* type is not an anomalous production but a systematic grammatical option in learner syntax. The learner data thus suggest that variation in prosody reflects variation in the structural resources selected in the numeration. The present analysis therefore supports the view that optionality in L2 grammar is not random but can be traced to identifiable points in the derivation.*

6.3 Implications for L2 grammar in general

The discussion above has several implications for the study of L2 grammar. First, and most fundamentally, the present data demonstrate that prosody provides us with keys to analyse learner syntax, especially those involving CP layers. A learner who produces a target-like sentence such as *Who do you believe likes Jack?* may be operating with a single-CP structure identical to that of native speakers, or with a two-CP structure involving an empty category in the embedded clause. Without prosodic evidence, these two possibilities cannot be distinguished. This means that previous studies which evaluated learner knowledge exclusively on the basis of surface word order may have underestimated the structural diversity of learner productions.

Second, the learner data support the view that L2 productions can reflect more than one derivational option. In the present study, this was observed both in target-like productions and in the *wh-wh* type. Some learner utterances were consistent with a structure containing a single CP, whereas others suggested a structure containing two CPs. This pattern indicates that learner grammar is not simply divided into target-like and non-target-like forms, but may involve multiple structural options underlying superficially similar outputs. The source of this optionality can be traced to the presence or absence of an embedded C head in the numeration, as proposed by Wakabayashi (1997) and Wakabayashi (2009).

Third, the results have implications for how non-target productions should be interpreted. Rather than treating them as random performance

* An anonymous reviewer pointed out that this is theoretically possible, but it is one of the possible analysis. We take this position at this study, but it may be true that learner creates two CPs but produce it with one IntP. We leave another possibility for further studies.

errors, the present analysis suggests that at least some of them reflect systematic grammatical possibilities in L2 syntax. In particular, the *wh-wh* type with two IntPs reflects a derivation consistent with Merge over Move and L1 transfer from Japanese *wh*-scope-marking constructions, while the *wh-wh* type with one IntP reflects a direct dependency derivation made possible by the absence of an embedded CP layer. In this sense, non-target productions may provide important evidence about the structural options available in interlanguage grammar.

Finally, the present study has methodological implications for L2 production research more broadly. If prosody can reveal structural differences that are not visible in surface form, then prosodic analysis should play a more central role in the investigation of L2 grammar. The findings of the present study support an approach in which syntax and prosody are examined together, rather than treated as independent components of learner performance. Future studies that adopt this approach may reveal further structural distinctions within production types that have previously been treated as uniform.

Our considerations in this study depend on the direct correlation between Intonational Phrase boundaries and CP structure. Our view on this transparent relationship between IntP and CP may appear to be too strong and require further verification, we regard this view is in the right line of consideration: We have no reason to propose any obstacle factors that obscure this direct mapping, i.e., when neither empirical evidence nor theoretical requirement applies, we should keep our assumption as simple as possible in this kind of scientific research, and so, our view is verified.

7. CONCLUSION

This study has examined the underlying syntactic structures of long-distance *wh*-questions in L2 English through the lens of prosody. The central claim is that prosody can serve as evidence for underlying syntax, and that target-like surface form does not necessarily imply target-like syntactic representation. The findings support this claim on multiple fronts and offer new insights into the structural diversity of interlanguage grammar.

The first main finding concerns target-like productions. The learner data showed that target-like utterances are not prosodically uniform. Some were realized as a single IntP, whereas others were realized as two IntPs. Under the prosody–syntax correspondence adopted in this study, this contrast suggests a structural difference: single-IntP productions are consistent with a single-CP structure, while two-IntP productions suggest a structure in which the embedded clause projects a separate CP containing an empty category. This finding directly challenges the assumption, widespread in GenSLA research, that target-like surface form implies target-like underlying syntax. The present data show that two learners who produce the same word order may be operating with structurally different grammars, a distinction that is invisible to analyses based on surface word order alone.

The second main finding concerns the *wh-wh* type. The learner data revealed that this production type also cannot be treated as a single uniform construction. Some *wh-wh* productions were realized with two IntPs, indicating a two-CP structure consistent with the indirect dependency approach to *wh*-scope-marking constructions. This pattern is compatible with the Merge over Move account proposed by Wakabayashi and Okawara (2003), in which the matrix *wh*-phrase is merged independently rather than derived by movement. In addition, the present study argues that L1 transfer from Japanese *wh*-scope-marking constructions plays a role in this pattern. The occurrence of *how* as a matrix scope marker in learner productions provides direct evidence for this claim, since this form corresponds to the Japanese scope marker *doo* (corresponding to English *how*) and cannot be attributed to Merge over Move alone.

The third main finding concerns the unexpected one-IntP *wh-wh* type. A subset of *wh-wh* productions was realized as a single IntP, a pattern that was not predicted under the initial hypothesis. This study proposes that these productions involve an embedded TP rather than an embedded CP. Because no embedded C head is present, the embedded *wh*-phrase is not valued within the embedded clause, and its [*uOp*] feature is free to move to the matrix clause under the direct dependency approach. This structural analysis explains both the one-IntP prosodic pattern and the difference between this type and the two-IntP *wh-wh* type. The contrast between the two *wh-wh* patterns can be reduced to a single structural choice: whether

the embedded clause projects a CP layer. The optionality of the embedded C head in the learner's numeration, as proposed by Wakabayashi (1997, 2009), accounts for the co-existence of both patterns within the same learner's data.

Taken together, these findings have two major implications. The first is theoretical. The present study demonstrates that interlanguage grammar is structurally more diverse than surface-form analyses suggest. Learner productions that appear identical on the surface may differ in their underlying representations, and this structural variation can be systematically revealed through prosodic evidence. The present findings therefore support the view that prosody is a reliable diagnostic tool for investigating L2 syntax, extending Royer (2022)'s proposal to the domain of second language acquisition.

The second implication is methodological. If prosody can reveal structural differences that are not visible in surface form, then production studies that rely solely on surface word order risk overlooking important dimensions of learner grammar. The present study suggests that a more complete picture of L2 syntactic knowledge requires that prosodic patterns be examined alongside surface form. This methodological point applies not only to the study of long-distance *wh*-questions but potentially to other constructions in which the relationship between surface form and underlying structure is not transparent.

Several issues remain for future research. First, the precise status of the empty category in target-like productions with two IntPs requires further investigation. Second, the present study was based on a relatively small number of participants, and the range of individual variation observed in the data suggests that a larger sample would be needed to establish more robust generalizations about learner grammar. Third, the role of prosody should be explored in a broader range of L2 constructions and learner populations, in order to assess the generality of the prosody-as-syntactic-evidence approach. Nevertheless, the present study has demonstrated that prosodic evidence offers an important window into the hidden structural variability of L2 grammar, and that individual variation in prosodic patterns reflects genuine structural differences in interlanguage syntax.

REFERENCES

- Anwyl-Irvine, Alexander L., Jessica Massonnié, Adam Flitton, Natasha Kirkham, and Jo K. Evershed. 2020. Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods* 52:388-407.
- Boersma, Paul, and David Weenink. 2024. *Praat: Doing phonetics by computer*, version 6.4.22. Computer program.
- Bošković, Željko. 1997. *The Syntax of Nonfinite Complementation: An Economy Approach*. Cambridge, MA: MIT Press.
- Bošković, Željko. 2007. On the locality and motivation of Move and Agree: An even more minimal theory. *Linguistic Inquiry* 38:589-644.
- Cable, Seth. 2010. *The Grammar of Q: Q-particles, Wh-movement, and Pied-piping*. Oxford: Oxford University Press.
- Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.
- Chomsky, Noam. 2000. Minimalist inquiries: The framework. In *Step by Step: Essays on Minimalist Syntax in Honor of Howard Lasnik*, eds. Roger Martin, David Michaels, Juan Uriagereka, and Samuel Jay Keyser, pp.89-155. Cambridge, MA: MIT Press.
- Chomsky, Noam. 2001. Derivation by phase. In *Ken Hale: A Life in Language*, ed. Michael Kenstowicz, pp.1-52. Cambridge, MA: MIT Press.
- Crain, Stephen, and Rosalind Thornton. 1998. *Investigations in Universal Grammar: A Guide to Experiments on the Acquisition of Syntax and Semantics*. Cambridge, MA: MIT Press.
- Crystal, Thomas H., and Arthur S. House. 1988. Segmental durations in connected-speech signals: Current results. *The Journal of the Acoustical Society of America* 83:1553-1573.
- Dayal, Veneta. 1993. Scope marking as indirect wh-dependency. *Natural Language Semantics* 2:137-170.
- Fujiwara, Yoshiki. 2021. Wh-scope marking in Japanese: Evidence for the indirect-dependency approach. *Linguistic Inquiry* 52:627-639.
- Hawkins, Roger, and Hajime Hattori. 2006. Interpretation of English multiple wh-questions by Japanese speakers: A missing uninterpretable feature account. *Second Language Research* 22:269-301.
- Kandybowicz, Jason. 2020. *Anti-contiguity: A Theory of Wh-prosody*. Oxford: Oxford University Press.
- Kimura, Takayuki. 2022. *Feature Selection, Feature Reassembly, and the Role of Universal Grammar: The Acquisition of Wh-questions by Japanese and Chinese Learners of English*. Hachioji: Chuo University dissertation.
- Kimura, Takayuki, and Shigenori Wakabayashi. 2024. UG-as-Guide in selection and reassembly of an uninterpretable feature in L2 acquisition of wh-questions: Evidence from islands and scope. In *Current Perspectives on Generative SLA: Processing, Influence, and Interfaces: Selected Proceedings of the 16th Generative Approaches to Second Language Acquisition Conference*, eds. Melita Velnić, Anne Dahl, and

- Kaja F. Listhaug, pp.318-350. Amsterdam: John Benjamins.
- Manetta, Emily. 2010. Wh-expletives in Hindi-Urdu: The vP phase. *Linguistic Inquiry* 41:1-34.
- McDaniel, Dana. 1989. Partial and multiple wh-movement. *Natural Language & Linguistic Theory* 7:565-604.
- Nespor, Marina, and Irene Vogel. 1986. *Prosodic Phonology*. Dordrecht: Foris.
- Royer, Justin. 2022. Prosody as syntactic evidence. *Natural Language & Linguistic Theory* 40:239-284.
- Saito, Mamoru. 1992. Long distance scrambling in Japanese. *Journal of East Asian Linguistics* 1:69-118.
- Schulz, Barbara. 2011. Syntactic creativity in second language English: Wh-scope marking in Japanese-English interlanguage. *Second Language Research* 27:313-341.
- Selkirk, Elisabeth. 2009. On clause and intonational phrase in Japanese: The syntactic grounding of prosodic constituent structure. *Gengo Kenkyu* 136:35-73.
- Selkirk, Elisabeth. 2011. The syntax-phonology interface. In *The Handbook of Phonological Theory*, 2nd edn., eds. John Goldsmith, Jason Riggle, and Alan C. L. Yu, pp.435-485. Malden, MA: Wiley-Blackwell.
- Wakabayashi, Shigenori. 1997. *The Acquisition of Functional Categories by Learners of English*. Cambridge: University of Cambridge dissertation.
- Wakabayashi, Shigenori. 2009. Lexical learning in second language acquisition: Optionality in the numeration. *Second Language Research* 25:335-341.
- Wakabayashi, Shigenori, and Izumi Okawara. 2003. Japanese learners' errors on long distance wh-questions. In *Generative Approaches to the Acquisition of English by Native Speakers of Japanese*, ed. Shigenori Wakabayashi, pp.215-245. Berlin: Mouton de Gruyter.
- Wakabayashi, Shigenori, and Yuki Takahashi. 2025. Variations underlying apparent tough constructions in English. In *Current Issues in L2 Teaching and Research*, eds. Stano Kong, Yu-Hsin Huang, Miao-Chin Chiu, Kai-Lin Wu, and One-Soon Her, pp.29-77. Taipei: Crane Publishing.
- Yamane, Maki. 2003. *On Interaction of First-Language Transfer and Universal Grammar in Adult Second Language Acquisition: Wh-movement in L1-Japanese/L2-English Interlanguage*. Storrs: University of Connecticut dissertation.

Tokitada Hangai
Chuo University
Tokyo, Japan
toki.hangai@gmail.com

Shigenori Wakabayashi
Chuo University
Tokyo, Japan
swkbys37@gmail.com

第二語言英語長距離*wh*疑問句的底層句法：來自韻律的洞見

半谷時忠、若林茂則
中央大学

本文以韻律為診斷工具，探討日本英語學習者產出長距離*wh*疑問句時的潛在句法結構。以往研究僅憑表面語序分析學習者產出，預設接近目標語的形式即反映目標語句法。本文採用韻律結構透明反映句法結構的觀點（Royer 2022），主張語調短語邊界可作為句法組織的診斷依據。誘發產出任務結果顯示，表面接近目標語的產出在韻律上並非均一：部分以單一語調短語實現，部分以兩個語調短語實現，反映出僅憑表面形式無法察覺的句法差異。*wh-wh*類型同樣呈現韻律變異，顯示至少存在兩種不同的潛在推導。本研究結果表明韻律分析可以作為探究第二語言語法的可靠診斷工具。

關鍵字：*wh*疑問句、英語、日語、句法韻律介面、變異

THE ROLE OF EXPLICITNESS OF L1 TENSE AND ASPECT MARKERS IN L2 TENSE ACQUISITION: EXAMINING THE CLINE OF DIFFICULTY HYPOTHESIS

Yuka Fujii
University of Tsukuba

ABSTRACT

This study examines how overt morphological realization and cross-linguistic form-meaning mappings influence learners' interpretation of tense-aspect meanings in L2 English. Drawing on insights from the Cline of Difficulty Hypothesis (COD; Cho and Slabakova 2014; Slabakova 2009) and the Feature Reassembly Hypothesis (FRH; Lardiere 2008, 2009), a Temporal Interpretation Choice Task was administered to Japanese and Chinese learners of English. Chinese learners, whose L1 lacks grammatical tense marking, achieved levels of interpretation accuracy comparable to those of Japanese learners. However, the two groups relied on different interpretive strategies: Japanese learners primarily depended on morphological cues, whereas Chinese learners relied more heavily on contextual information. In addition, responses became more variable when direct L1-L2 form-meaning correspondences were unavailable. These findings suggest that explicit morphological realization alone does not fully account for learners' tense-aspect interpretation. Rather, successful interpretation appears to be shaped by the interaction of cue availability, cross-linguistic form-meaning mappings, and learners' strategies for resolving competing interpretations.

Keywords: Cline of Difficulty Hypothesis; Cross-Linguistic Form-Meaning Mapping; Explicitness of Tense-Aspect Markers; L1 Japanese and Chinese Learners of English; Tense-Aspect Interpretation

1. INTRODUCTION

In second language acquisition research, a number of hypotheses have been proposed to predict learning difficulties (e.g., Interface Hypothesis (Sorace 2011), Interpretability Hypothesis (Tsimpli and Dimitrakopoulou 2007), Feature Reassembly Hypothesis (FRH; Lardiere 2008, 2009)). Among these, the FRH has provided a particularly influential framework, as it offers a principled account of how learners map and reassemble linguistic features across languages. However, while the FRH specifies the process of acquisition, it remains less clear what determines the degree of difficulty involved in this process.

One attempt to address this issue is the Cline of Difficulty Hypothesis (Cho and Slabakova 2014; Slabakova 2009), grounded in the FRH, which predicts that features that are more explicit are easier to acquire than covert features that rely on discourse tracking. This account attributes acquisition difficulty primarily to the explicitness of formal realization. However, the proposal has been primarily examined in the domain of article acquisition, leaving open the question of its generalizability to other domains. At the same time, recent work has questioned the role of explicitness in determining acquisition difficulty. Kume (2021) suggests that difficulty may not primarily stem from the degree of overt marking, but rather from the complexity of feature reassembly in cross-linguistic form-meaning mappings across languages. These contrasting perspectives raise the question of how explicit morphological realization and cross-linguistic form-meaning mappings jointly shape learners' interpretation of tense-aspect meanings.

The present study investigates whether explicitness plays a decisive role in tense-aspect acquisition by comparing learners with different L1 backgrounds. Specifically, it examines Japanese learners of English (JLEs), whose L1 encodes temporal distinctions through overt morphological marking, and Chinese learners of English (CLEs), whose L1 does not primarily encode temporal reference through grammatical tense morphology. By directly comparing these groups, the study aims to examine whether learners can achieve comparable interpretive outcomes through different cue-reliance strategies highlighting the role of competing form-meaning mappings and the facilitating function of

explicitness. In doing so, the study re-examines the role of explicit morphological realization in tense-aspect interpretation while highlighting the contribution of cross-linguistic form-meaning mappings and cue-reliance strategies.

The remainder of this paper is organized as follows. Section 2 reviews the Cline of Difficulty and the Feature Reassembly Hypothesis, and previous studies of tense and aspect markers acquisition. Section 3 provides an overview of tense and aspect marking in the relevant languages. Section 4 presents the experimental design and results, followed by a discussion in Section 5.

2. THEORETICAL BACKGROUND

2.1 Feature Reassembly Hypothesis vs. Cline of Difficulty Hypothesis

The Feature Reassembly Hypothesis (FRH; Lardiere 2008, 2009) proposes that the acquisition difficulty arises from the *mapping* and *reassembling* of features across languages. At early stages of acquisition, learners initially map L2 forms onto L1 representations based on semantic similarity. However, when no direct correspondence exists, they must reassemble features into new configurations. Difficulty is therefore predicted to arise when form-meaning mappings differ across languages, particularly in cases involving complex correspondences, where multiple competing mappings must be resolved.

Building on this perspective, the Cline of Difficulty Hypothesis (COD; Cho and Slabakova 2014; Slabakova 2009) proposes that overtly realized features are easier to acquire than covert features that rely on discourse-level computation. According to this account, the overt and covert realization of linguistic features is expected to influence acquisition difficulty.

Thus, whereas the FRH highlights the role of form-meaning mapping complexity in acquisition, the COD specifies a particular source of difficulty related to the explicitness of feature realization. These

perspectives are not necessarily mutually exclusive; rather, they emphasize different factors that may contribute to acquisition outcomes.

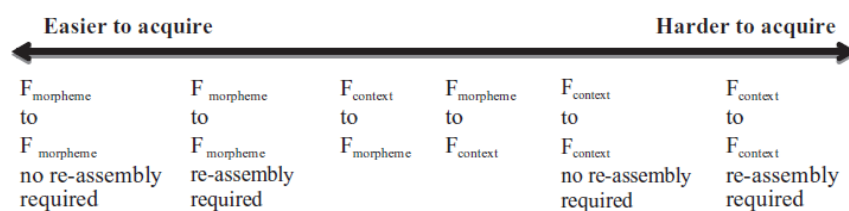


Figure 1. Model of Cline of Difficulty (Adopted from Cho and Slabakova 2014; 166)

While the FRH has been widely applied across domains as a framework for explaining acquisition difficulty, empirical support for the COD has largely been restricted to the domain of article acquisition. In addition, recent work has questioned whether explicitness alone can account for acquisition difficulty. Kume (2021), investigating the definiteness acquisition by Korean and English learners of Japanese, argues that acquisition difficulty is more closely tied to the complexity of reassembling features than to the degree of overt marking.

This suggestion raises the possibility that explicitness alone may not fully account for learners' interpretation of tense-aspect meanings, highlighting the need to re-examine the role of explicitness alongside cross-linguistic form-meaning mappings in tense-aspect interpretation.

2.2 Previous Research on Tense and Aspect Acquisition

Previous studies on tense-aspect acquisition provided mixed support for these accounts. Muroya (2018) showed that Japanese learners of English exhibit higher accuracy for morphemes that have clear counterparts in their L1 (e.g., English past tense *-ed* corresponding to Japanese *-ta*), while showing difficulty with forms lacking direct equivalents (e.g., third-person suffix *-s*). This pattern is consistent with the role of L1-based mapping in shaping acquisition outcomes. In contrast, Slabakova (2015) found that English learners of Chinese, whose L1 has overt tense markers and L2 does not, were able to interpret temporal

reference accurately, relying instead on contextual and grammatical cues. This finding suggests that learners can successfully compensate for the lack of explicit morphology by relying on contextual and grammatical cues.

Despite these insights, important gaps remain. First, previous research has largely focused on one-directional acquisition (e.g., English to Chinese), leaving underexplored how learners with a tenseless L1 acquire tense in a language with overt tense marking. Second, few studies have directly compared learners with different L1 backgrounds within the same experimental paradigm, limiting our understanding of the generalizability of these theoretical claims.

The present study addresses these gaps by comparing Japanese and Chinese learners of English in their interpretation of tense-aspect morphology. In doing so, it explores the relative contribution of overt morphological realization, as emphasized by the COD, and cross-linguistic form-meaning mapping, as highlighted in the FRH, to learner's interpretation of tense-and aspect meanings.

3. LINGUISTIC BACKGROUND

This section provides an overview of tense and aspect marking in Japanese, Chinese and English. Throughout this paper, the term Chinese is used to refer specifically to Mandarin Chinese.

The following comparison is intended to illustrate representative form-meaning correspondences that learners may encounter when mapping tense-aspect meanings across languages. It is not intended to imply complete semantic equivalence between the forms listed. Rather, the comparison focuses on functional similarities that are relevant to L2 acquisition and cross-linguistic mapping.

3.1 Tense in Japanese, Chinese, and English

Tense is generally defined as a grammatical category that locates a situation in time and expresses the temporal relationship between the reference time and the event through morphological or syntactic cues (Shirai and Andersen 1995). English distinguishes between present and

past tense, whereas Japanese contrasts non-past and past forms. In contrast, Chinese is commonly described as lacking a grammatical category of tense (Liu 2015; Smith and Erbaugh 2005). This section reviews how temporal reference is encoded in these three languages.

3.1.1 Present tense

In Japanese, the *-ru* form (the base form of verb ending in *-u*) is used to indicate non-past, typically present-time reference (Iori and Tagawa 2019). In Chinese and English, the base form of verb represents present-time reference, with English additionally marking third-person singular subjects with the suffix *-s* (Muroya 2018; Nishizaka 2019). The following examples illustrate these differences:

(1) Japanese

watashi-wa	koen-de	aru-ku.
I-NOM	park-LOC	walk-PRES

“I walk in the park.”

(Based on Nishizaka (2019); constructed by the author)

(2) Chinese

wo	zai	gongyuan	zou.
I	LOC	park	walk-PRES

“I walk in the park.”

(Based on Nishizaka (2019); constructed by the author)

(3) English

I walk in the park.

(4) English (third-person singular subject)

He walks in the park.

(Based on Muroya (2018); slightly modified by the author)

3.1.2 Past tense

The expression of the past time differs substantially across the three languages. In Japanese, past tense is marked by the suffix *-ta*, which replaces the final *-u* of the base verb form (Muroya 2018). In English, past tense is typically indicated by the addition of the suffix *-ed* to the base

form of verb or using irregular forms (Muroya 2018). In contrast, Chinese does not employ verbal inflection and lacks overt tense marking. Instead, past-time reference is conveyed through temporal adverbs, aspect markers, and contextual cues (Smith and Erbaugh 2005). The following examples illustrate these differences:

(5) Japanese

watashi-wa	koen-de	aru-ita.
I-NOM	park-LOC	walk-PAST

“I walked in the park.”

(Based on Nishizaka (2019); constructed by the author)

(6) Chinese

wo	zuotian	zai	gongyuan	zou.
I	yesterday	LOC	park	walk

“I walked in the park yesterday.”

(Based on Nishizaka (2019); constructed by the author)

(7) English

I walked in the park.

(8) English (irregular form)

I ran in the park.

(Based on Muroya (2018); slightly modified by the author)

These examples illustrate that English and Japanese employ morphological markers to encode past time, whereas Chinese relies on other linguistic cues. Accordingly, the expression of tense differs systematically across the three languages. For the purposes of this paper, morphological elements used to encode tense will be referred to as “tense markers.”

3.2 Grammatical Aspect in Japanese, Chinese, and English

Aspect refers to the temporal properties of an event, highlighting the internal temporal structure, such as whether ongoing or completed. This section focuses on the progressive, resultative, and perfective aspect as realized in these three languages.

3.2.1 Progressive aspect

Progressive aspect denotes an ongoing event. In Japanese, it is formed by attaching *-tei-ru* to the verb stem. In Chinese, it is expressed by putting the particles *-zai* or *-zhe* immediately before or after the base form of an activity verb. In English, it is expressed by placing the auxiliary *be* immediately before verb and adding the *-ing* to the base form of verb (Liu 2015; Muroya 2018; Nishizaka 2019; Zhang 2001). The following examples illustrate these differences:

(9) Japanese

watashi-wa	koen-de	arui-tei-ru.
I-NOM	park-LOC	walk-PROG-PRES

“I am walking in the park.”

(Based on Nishizaka (2019); constructed by the author)

(10) Chinese (*-zai-*)

wo	zai	gongyuan	zai-zou.
I	LOC	park	PROG-walk

“I am walking in the park.”

(11) Chinese (*-zhe*)

wo	zai	gongyuan	zou-zhe.
I	LOC	park	walk-PROG

“I am walking in the park.”

(Based on Nishizaka (2019); constructed by the author)

(12) English

I am walking in the park.

(Based on Muroya (2018); slightly modified by the author)

3.2.2 Resultative aspect

Resultative aspect denotes a state resulting from a completed event. In Japanese, *-tei-ru* is also used, particularly with accomplishment verbs and depending on the context (Nishizaka 2019). In Chinese, it is indicated by *-le* or *-zhe* immediately after the base form of verb, with *-zhe* typically restricted to accomplishment verbs (Inoue 2019; Liu 2015; Zhang 2001).

In English, it is typically expressed using the auxiliary *have* with the suffix *-ed* (Muroya 2018). The following examples illustrate these differences:

- (13) Japanese
sakura-no hana-ga sai-tei-ru.
cherry blossom-GEN flower-LOC bloom-RESL-PRES
“The cherry blossoms have bloomed.”
(Zhang, 2001:140)
- (14) Chinese (-*le*)
wo jiehun le.
I marry RESL
“I got married.”
- (15) Chinese (-*zhe*)
yinghuarshu shang yinghuar shengkai-zhe.
cherry blossom LOC the flower- bloom-RESL
tree LOC
“The cherry blossoms have bloomed.”
(Zhang, 2001:140)
- (16) English
The cherry blossoms have bloomed.
(Based on Muroya (2018); constructed by the author)

3.2.3 Perfective aspect

Lastly, perfective aspect denotes completed events without necessarily highlighting the result state. In Japanese, it is expressed by *-ta*, which also functions as a past tense marker (Nishizaka 2019). In Chinese, it is indicated by *-le* immediately after the base form of activity verb. In English, it is expressed through the auxiliary *have* with the suffix *-ed*. Following Muroya (2018), the present study adopts this classification for purposes of cross-linguistic comparison. This classification is intended to highlight representative functional correspondences that learners may draw upon when mapping temporal and aspectual meanings across languages. It is not intended to imply complete semantic equivalence among the forms discussed. The following examples illustrate these differences:

- (17) Japanese
 watashi-wa koen-de aru-ita.
 I-NOM park-LOC walk-PERF
 “I have walked in the park.”
 (Based on Nishizaka (2019); constructed by the author)
- (18) Chinese
 wo zai gongyuan zou-le.
 I LOC park walk-PERF
 “I have walked in the park.”
 (Based on Nishizaka (2019); constructed by the author)
- (19) English
 I have walked in the park.
 (Based on Muroya (2018); constructed by the author)

3.3 The Complexity of Form-Meaning Mapping

As illustrated above, cross-linguistic mappings between Japanese, Chinese, and English tense and aspectual markers are complex. A single marker may correspond to multiple meanings, and the semantic range of these markers differs across languages. This many-to-many mapping creates potential difficulty for learners, particularly when acquiring English tense-aspect morphology from L1 Japanese or Chinese. While the present overview is not exhaustive, Table 1 summarizes the relevant correspondences examined in this study.

Importantly, the nature of this mapping complexity differs across learner groups. For Japanese learners of English (JLEs), whose L1 encodes tense distinctions through overt morphological marking, mapping to English tense-aspect forms may be relatively straightforward. In contrast, Chinese learners of English (CLEs) are expected to rely more heavily on aspectual and contextual cues because temporal reference in Chinese is not primarily encoded through grammatical tense morphology. As a result, learners may need to establish correspondences between multiple cue sources and English tense-aspect forms, potentially increasing the complexity of cross-linguistic mapping.

Table 1. Correspondence between Japanese, Chinese and English Tense and Aspect Markers (Based on Muroya (2018); and Chinese part was constructed by the author)

		Japanese	Chinese		English
Tense	Present	<i>-(r)u</i>	null		null / <i>-s</i>
	Past	<i>-ta</i>	null		<i>-ed</i>
Aspect	Perfective		<i>-le</i> (了)	<i>-guo</i> (过)	have <i>-ed</i>
	Resultative			<i>-zhe</i> (着)	
	Progressive	<i>-tei-ru / ta</i>	<i>zai-</i> (在)		be <i>-ing</i>

From the perspective of the COD, different predictions can be made for these learner groups. JLEs, whose L1 encodes tense through overt morphological marking, are expected to fall at the lowest level of difficulty in the model (see Table 1 above). In contrast, CLEs, whose L1 lacks overt tense marking, are predicted to face greater difficulty. However, Slabakova (2015) showed that English learners of Chinese were able to successfully interpret temporal reference in both judgment and production tasks despite the absence of overt tense marking, a condition that corresponds to an even higher level of difficulty in the model. It suggests that learners can compensate for the lack of explicit morphology by relying on alternative cues. This raises the question of whether explicitness alone can account for acquisition difficulty adequately.

The present study addresses this issue by examining the relative contributions of explicit morphological realization and form-meaning mapping complexity to learners' interpretation of English tense-aspect meanings. The Research Questions in this research are shown below.

RQ1: Do learners rely more on morphologic cues or contextual temporal cues, and does this differ across L1 backgrounds?

RQ2: Does the degree of alignment between L1 and L2 form-meaning mappings affect learners' interpretation preferences?

4. EXPERIMENT

This study employed the Temporal Interpretation Choice Task (TICT) to examine how Japanese learners of English (JLEs) and Chinese learners of English (CLEs) interpret tense, drawing on their L1-based strategies (cf. Slabakova 2015).

4.1 Participants

The participants were 24 JLEs and 21 CLEs. The JLE group consisted of undergraduate students with English proficiency in the B2–C1 range of the CEFR, as indexed by TOEIC L&R IP test. The CLE group consisted of graduate students whose proficiency levels were comparable to those of the JLEs, except for one participant at the B1 level and another at the C2 level. For the CLEs, proficiency levels were estimated by converting scores from standardized English tests (e.g., IELTS and TOEFL iBT) into TOEIC L&R equivalents using established conversion guidelines. With respect to English learning experience, all participants had received more than eight years of formal English instruction in their respective countries. None had resided in an English-speaking country for more than one month, except for one CLE participant who had stayed in the United Kingdom for one year as an exchange student. Thus, despite the difference in academic status, the two groups had broadly comparable English learning backgrounds.

4.2 Materials

The materials consisted of a context sentence, one target sentence, and the answer options as an item. The context sentence included a tense marker (present or past) and a temporal adverb (e.g., *now*, *yesterday*) in the participants' L1. Temporal adverbs were intentionally included to establish a clear temporal frame and to investigate how learners integrate contextual and morphological information during tense-aspect interpretation. This manipulation was designed to reveal potential differences in cue reliance across learner groups.

The target sentence, also presented in the L1, contained a tense marker (present or past) and an aspect marker (simple, progressive, and perfective). The answer options comprised four English verb forms

reflecting a 2×2 design crossing tense (present vs. past) and aspect (simple vs. progressive, simple vs. perfective) (e.g., *takes/is taking/took/was taking*). For the Chinese materials, only aspect markers were manipulated in the target sentences, as temporal reference in Chinese is primarily determined by contextual cues.

The presentation of the context and target sentences in the participants' L1 was intentional because the study aimed to examine cross-linguistic form-meaning mapping and the extent to which learners rely on L1-based interpretive strategies when processing L2 tense-aspect meanings.

In the VP condition, only activity verbs were used in the target sentences to minimize the influence of lexical aspect. Most of the VP items were adopted from the TICT materials of Slabakova (2015), with each condition comprising four tokens, and the context sentences were constructed by the author.

Table 2. Material of the TICT for Japanese Learners of English
(Upper: Context sentence; Lower: Target sentence)

Condition (Context tense / Target tense)	Target aspect	
	Simple (- <i>(r)u</i> / - <i>ta</i>)	Progressive (- <i>tei-ru</i> / - <i>ta</i>)
Present / Present	現在張三は家にいる。 彼は風呂に入る。 (Zhangsan is at home now. He takes a bath.)	現在張三は家にいる。 彼は風呂に入っている。 (Zhangsan is at home now. He is taking a bath.)
Present / Past	現在張三は家にいる。 彼は風呂に入った。 (Zhangsan is at home now. He took a bath. / He has taken a bath.)	現在張三は家にいる。 彼は風呂に入っていた。 (Zhangsan is at home now. He was taking a bath.)
Past / Present	昨日張三は買い物に行った。 彼は車を運転する。 (Zhangsan went shopping yesterday. He drives a car.)	昨日張三は買い物に行った。 彼は車を運転している。 (Zhangsan went shopping yesterday. He is driving a car.)
Past / Past	昨日張三は買い物に行った。 彼は車を運転した。 (Zhangsan went shopping yesterday. He drove a car.)	昨日張三は買い物に行った。 彼は車を運転していた。 (Zhangsan went shopping yesterday. He was driving a car.)

Table 3. Material of the TICT for Chinese Learners of English
(Upper: Context sentence; Lower: Target sentence)

Condition (Context tense)	Target aspect			
	Simple (null)	Progressive (<i>zai-</i>)	progressive (<i>-zhe</i>)	perfective (<i>-le</i>)
Present	现在李四在家。 他洗澡。 (Lisi is at home now. He takes a bath.)	现在李四在家。 他 在 洗澡。 (Lisi is at home now. He is taking a bath.)	现在李四在家。 他洗 着 澡。 (Lisi is at home now. He is taking a bath.)	现在李四在家。 他洗澡 了 。 (Lisi is at home now. He took a bath. / He has taken a bath.)
Past	昨天张三在外 面。 他开车。 (Zhangsan was outside yesterday. He drove a car.)	昨天张三在外 面。 他 在 开车。 (Zhangsan was outside yesterday. He was driving a car.)	昨天张三在外 面。 他开 着 车。 (Zhangsan was outside yesterday. He was driving a car.)	昨天张三在外 面。 他开车 了 。 (Zhangsan was outside yesterday. He drove a car.)

4.3 Design

The materials were divided into two lists following a pseudo-Latin square design, with each list containing two items per condition (e.g., 16 target items per list). In addition, 40 filler items were included to reduce potential response biases and discourage participants from developing task-specific strategies.

4.4 Procedure

The TICT was conducted via Google Forms. A context sentence, a target sentence, and the answer options were presented within the same screen. As shown in Figure 2, participants read a context sentence, followed by a target sentence, and selected one of four answer options (see Section 4.2) to indicate the temporal interpretation of the target sentence. The sentences remained visible during response selection, and participants were allowed to revise their responses within a trial but not permitted to return to previous items. All items were presented in a randomized order.

The participants completed the TICT and a background questionnaire within 30 minutes.



現在李四は家でくつろいでいる。彼はお風呂に入っている。*

☐ takes a bath
☐ took a bath
☐ is taking a bath
☐ was taking a bath

Answer options

Figure 2. Screenshot of the Temporal Interpretation Choice Task

4.5 Data Analysis

Response distributions were analyzed descriptively. For each condition, response proportions were computed separately for each group, with the total number of responses per condition (48 for JLEs and 42 for CLEs) treated as 100%. The distribution of the four answer options was then examined to characterize tendencies in temporal interpretation across conditions and groups. In Table 4 and 5, percentages are reported relative to number of the total responses, with raw response counts provided in parentheses.

In addition to the descriptive analyses, generalized linear mixed-effects models (GLMMs) with a binomial link function were conducted to examine whether learners' temporal interpretations were systematically affected by contextual tense and tense-aspect markers. For JLEs, context, tense marker, aspect marker, and proficiency were included as fixed effects, with participant and item as random intercepts. For CLEs, contextual tense, marker type, and proficiency were included as fixed effects, with participant as a random intercept. Estimated marginal means (EMMeans) were calculated to facilitate interpretation of significant effects.

4.6 Result

As expected, JLEs showed a strong preference for English tense-aspect forms that align with their L1 counterparts. This pattern was observed across conditions, including those in which the context sentence and the target sentence were mismatched, suggesting a reliance on L1-based form-meaning mappings rather than on contextual temporal cues, as can be seen in the comparisons between gray-shaded cells and green-shaded cells in Table 4.

Table 4. Response Rate of Japanese Learners of English
(Bold: The most selected option; Gray-shaded cells; Preferable option)

Context	Marker	Pres_ Simple (takes)	Pres_ Progre ssive (is taking)	Past_ Simple (took)	Past_ Progre ssive (was taking)	Pres_ Perfe ctive (has taken)	Past_ Perfec tive (had taken)
Present	-(r)u	79.17 (38)	16.67 (8)	2.08 (1)	2.08 (1)	-	-
Present	-ta	27.08 (13)	-	50.00 (24)	-	22.92 (11)	0.00 (0)
Present	-tei-ru	12.60 (6)	85.42 (41)	0.00 (0)	2.08 (1)	-	-
Present	-tei-ta	12.60 (6)	10.42 (5)	6.25 (3)	70.83 (34)	-	-
Past	-(r)u	50.00 (24)	6.25 (3)	29.17 (14)	14.58 (7)	-	-
Past	-ta	2.08 (1)	-	87.50 (42)	-	4.17 (2)	6.25 (3)
Past	-teiru	4.17 (2)	62.5 (30)	8.33 (4)	25.00 (12)	-	-
Past	-tei-ta	4.17 (2)	2.08 (1)	10.42 (5)	83.33 (40)	-	-

With respect to the accuracy of the temporal interpretation, JLEs generally selected the preferable responses across conditions, indicating

that they were largely able to identify the intended temporal interpretation. However, variability in responses was observed in two contexts.

First, when a marker was compatible with multiple interpretations (e.g., *-ta*, *-tei-ru*), responses tended to be distributed across competing options, reflecting the complexity arising from one-to-many form-meaning mapping. At the same time, however, clear response preferences were still observed in the present data. In target sentences with the tense-aspect marker *-ta*, JLEs consistently favored past tense interpretation over perfective or resultative readings, even though multiple interpretations are acceptable. Similarly, responses to the aspect marker *-tei-ru* were strongly skewed toward progressive interpretations, with resultative readings rarely selected. This tendency may be attributed to the exclusive use of activity verbs in the present study, which reduce the availability of resultative readings.

Second, increased variability was observed in context-target mismatch conditions. In such cases, the temporal relation between the context and the target sentence may have been less readily integrated. However, JLEs tended to rely on overt tense morphology when interpreting temporal relations, even when contextual cues were in conflict.

To evaluate these descriptive tendencies statistically, GLMM was conducted. The analysis revealed a significant interaction between context and tense marker ($\beta = 3.89, z = 5.10, p < .001$). Estimated marginal means showed that interpretation accuracy was highest when context and tense marker were aligned (Past–Past = .87; Present–Present = .84), whereas accuracy decreased in mismatch conditions (Present–Past = .62; Past–Present = .25). Pairwise comparisons showed no significant difference between the two aligned conditions ($p = .936$). In contrast, the Past–Present condition yielded significantly lower accuracy than both aligned conditions ($ps < .001$), and the Present–Past condition also showed significantly lower accuracy than the aligned conditions ($ps < .01$). These results indicate that JLEs were strongly influenced by the alignment between contextual information and overt tense morphology.

In contrast to JLEs, CLEs appeared to rely more on the context information when interpreting the target sentence. This pattern suggests that CLEs made use of contextual temporal cues rather than relying primarily on aspect markers in determining temporal reference.

Surprisingly, despite the absence of overt tense marker in Chinese, which might be expected to increase matching difficulty, CLEs showed accuracy levels comparable to JLEs.

Interestingly, in the null condition, where no overt aspectual marking was provided, CLEs tended to select progressive interpretations, as can be seen in the comparison between gray-shaded cells and pink-shaded cells in Table 5. This pattern suggests that learners may tend toward an event-in-progress interpretation when constructing present-time reference, or that their interpretations may be influenced by the meaning of temporal adverb (e.g., *now*) in the context sentence.

Table 5. Response Rate of Chinese Learners of English
(Bold: The most selected option; Gray-shaded cells; Preferable option)

Context	Marker	Pres_ Simple (takes)	Pres_ Progressive (is taking)	Past_ Simple (took)	Past_ Progressive (was taking)
Pres	null	21.43 (9)	78.57 (33)	0.00 (0)	0.00 (0)
Pres	<i>zai-</i>	2.38 (1)	97.62 (41)	0.00 (0)	0.00 (0)
Pres	<i>-zhe</i>	0.00 (0)	97.62 (41)	0.00 (0)	2.38 (1)
Pres	<i>-le</i>	7.14 (3)	0.00 (0)	73.81 (31)	19.05 (8)
Past	null	11.90 (5)	2.38 (1)	61.90 (26)	23.81 (10)
Past	<i>zai-</i>	4.76 (2)	2.38 (1)	7.14 (3)	95.71 (36)
Past	<i>-zhe</i>	4.76 (2)	2.38 (1)	4.76 (2)	88.10 (37)
Past	<i>-le</i>	4.76 (2)	0.00 (0)	73.81 (31)	21.43 (9)

The descriptive patterns were further examined using GLMM. The analysis revealed a significant effect of context ($\beta = 0.90$, $z = 1.98$, p

= .048), indicating higher overall accuracy in present contexts (.78) than in past contexts (.63). Significant effects were also observed for marker. The progressive markers *zai-* (.94) and *-zhe* (.95) yielded the highest estimated accuracy, whereas the null condition (.40) showed substantially lower accuracy. Furthermore, a significant interaction was observed between present context and the null marker ($\beta = -2.74$, $z = -4.04$, $p < .001$). Estimated marginal means indicated that accuracy in the null condition dropped from .62 in past contexts to .21 in present contexts. These findings suggest that CLEs relied on both contextual and aspectual cues, with contextual information becoming particularly important when overt aspectual marking was absent.

5. DISCUSSION

The present study examined how L2 learners map form and meaning in tense-aspect interpretation, focusing on the role of cross-linguistic alignment between L1 and L2 representations.

With respect to RQ1, the results showed that learners differed in how they weighted available sources of information during tense-aspect interpretation. JLEs primarily relied on L1-based morphological cues, whereas CLEs drew on a combination of contextual and aspectual cues. Despite these differences in cue weighting, the two groups achieved comparable levels of interpretation accuracy, indicating that learners can arrive at similar interpretive outcomes through different cue-reliance strategies.

One possible explanation is that some CLEs may partially reinterpret certain aspect markers in tense-like terms. In particular, the perfective marker *-le* is associated with past-time reference and may therefore be treated by some learners as a marker of past tense, despite its aspectual status in Chinese. If so, some responses may reflect simplified tense-based representations rather than full sensitivity to aspectual distinctions. Nevertheless, the comparable levels of accuracy observed across learner groups suggest that successful tense-aspect interpretation cannot be explained solely by the presence or absence of overt tense morphology.

Rather, learners appear able to compensate for the lack of overt tense marking through the use of contextual and aspectual information.

From the perspective of the Cline of Difficulty Hypothesis (COD), these findings suggest that explicit morphological realization may influence the types of cues learners rely on during interpretation. JLEs showed a stronger tendency to rely on morphological information, whereas CLEs relied more heavily on contextual cues. However, because both groups ultimately achieved comparable levels of interpretation accuracy, explicitness alone does not appear sufficient to account for learners' performance in tense-aspect interpretation.

With respect to RQ2, the degree of alignment between L1 and L2 form-meaning mappings had a clear effect on learners' interpretation. In aligned conditions, learners tended to select responses consistent with L1-based mappings, indicating facilitative effects of cross-linguistic similarity. In contrast, in mismatched conditions, responses became more variable, suggesting increased difficulty when direct correspondence was not available. This pattern indicates that learners must rely on multiple sources of information when L1-based mappings cannot be straightforwardly applied.

These findings are consistent with the Feature Reassembly Hypothesis (FRH), which emphasizes the role of cross-linguistic form-meaning mappings in shaping learner behavior. In particular, the increased variability observed in mismatched conditions can be understood as reflecting the difficulty of resolving competing mappings across languages. This finding highlights the importance of mapping complexity in learners' interpretation of tense-aspect meanings.

However, several limitations of the present study should be acknowledged. First, the relatively small number of participants and items per condition may limit the generalizability of the findings. Future research should include a larger number of participants and items to provide a more robust test of the hypotheses examined here. Therefore, the present findings should be interpreted as preliminary evidence suggesting that both cross-linguistic form-meaning mapping and the availability of morphological and contextual cues contribute to learners' interpretation of tense-aspect meanings.

Second, English proficiency was assessed using different instruments across the two learner groups. Even if the two groups showed broadly comparable proficiency profiles, the use of different assessment instruments may have introduced additional variability. Therefore, comparisons between the two learner groups should be interpreted with caution.

Third, despite efforts to restrict the stimuli to activity verbs, subtle differences in lexical semantic properties may have influenced participants' interpretations. Future studies should employ a more tightly controlled set of lexical items to further isolate the effects of tense-aspect morphology, while also considering potential cross-linguistic differences in lexical semantic properties across languages.

Lastly, although this design was necessary to investigate cross-linguistic mapping, it may also have increased the salience of L1-based strategies. Future research employing L2-only materials would help determine the extent to which the present findings generalize to more naturalistic processing contexts.

In summary, the present findings suggest that explicit morphological realization represents only one factor influencing learners' interpretation of tense-aspect meanings. While overt morphology may affect the types of cues learners preferentially rely on, successful interpretation also depends on the availability of contextual information and the degree of correspondence between L1 and L2 form-meaning mappings. A fuller account of tense-aspect interpretation should therefore consider the interaction among cue availability, mapping complexity, and learners' strategies for resolving competing interpretations.

REFERENCES

- Cho, Jacee, and Roumyana Slabakova. 2014. Interpreting definiteness in a second language without articles: The case of L2 Russian. *Second Language Research* 30(2):159-190. <https://doi.org/10.1177/0267658313509647>
- Inoue, Masaru. 2019. Chugokugo-no "suru" to "shita" to "shiteiru" 中国語の「する」と「した」と「している」 [Chinese "do", "did", and "be doing"]. In *Nihongo-no tense-aspect kenkyu-wo toinaosu, dai 1 kan: "suru" no sekai* 日本語のテンス・アスペクト研究を問い直す 第1巻: 「する」の世界

- [Reconsidering tense and aspect in Japanese Vol.1: the world of “suru”], eds. Isao Iori, and Takumi Tagawa, pp.141-161. Tokyo: Hitsuji shobou.
- Iori, Isao, and Takumi Tagawa. 2019. *Nihongo no tense-aspect kenkyuu wo toinaosu dai 1 kan: “suru” no sekai* 日本語のテンス・アспект研究を問い直す 第1巻: 「する」の世界 [Reconsidering tense and aspect in Japanese Vol.1: the world of “suru”]. Tokyo: Hitsuji shobou.
- Kume, Keisuke. 2021. *The acquisition of definiteness expressions in non-native Japanese by English- and Korean-speaking learners*. Doctoral Dissertation, University of York, UK.
https://etheses.whiterose.ac.uk/29098/1/Kume_204039692_Thesis.pdf
- Lardiere, Donna. 2008. Feature assembly in second language acquisition. In *The Role of Formal Features in Second Language Acquisition*, eds. Juana M. Liceras, Helmut Zobl, and Helen Goodluck, pp.106-40. New York: Routledge.
<https://doi.org/10.4324/9781315085340>
- Lardiere, Donna. 2009. Some thoughts on the contrastive analysis of features in second language acquisition. *Second Language Research* 25:173-227.
<https://doi.org/10.1177/0267658308100283>
- Liu, Meichun. 2015. Tense and aspect in Mandarin Chinese. In *The Oxford Handbook of Chinese Linguistics*, eds. William S-Y. Wang, and Chaofen Sun, pp.274-289. Oxford University Press.
<https://doi.org/10.1093/oxfordhb/9780199856336.013.0070>
- Muroya, Akiko. 2018. L1 Transfer in L2 Acquisition of English Verbal Morphology by Japanese Young Instructed Learners. *Languages* 4(1):1.
<https://doi.org/10.3390/languages4010001>
- Nishizaka, Shohei. 2019. *Tense-Aspect Acquisition of Japanese as a Second Language by Chinese Speakers*. Doctoral Dissertation, Nagoya University, Japan.
<https://nagoya.repo.nii.ac.jp/records/28596>
- Shirai, Yasuhiro, and Roger W. Andersen. 1995. The acquisition of tense-aspect morphology: a prototype account. *Language* 71(4): 743-762.
<https://doi.org/10.2307/415743>
- Slabakova, Roumyana. 2015. Acquiring Temporal Meanings Without Tense Morphology: The Case of L2 Mandarin Chinese. *The Modern Language Journal*, 99 (2):283-307. <http://www.jstor.org/stable/43650028>
- Smith, Carlota S., and Mary S. Erbaugh. 2005. Temporal interpretation in Mandarin Chinese. *Linguistics*, 43 (4): 713-756. <https://doi.org/10.1515/ling.2005.43.4.713>
- Sorace, Antonella. 2011. Pinning down the concept of “interface” in bilingualism. *Linguistic Approaches to Bilingualism* 1(1):1-33.
<https://doi.org/10.1075/lab.1.1.01sor>
- Tsimpli, Ianthi Maria and Maria Dimitrakopoulou, 2007. The Interpretability Hypothesis: Evidence from wh-interrogatives in Second Language Acquisition. *Second Language Research* 23 (2):215-242.
<https://doi.org/10.1177/0267658307076546>

Zhang, Linsheng. 2001. *Nihongo kyoiku no tame-no goyou bunnseki: chuugokugo washa-no bogo kanshou 20 rei* [Misuse Analysis for Japanese Language Teaching: 20 Examples of Native Language Interference by Chinese Speakers]. 3A network.

Yuka Fujii
Degree Program in Humanities and Social Sciences
University of Tsukuba
Ibaraki, Japan
yukaaa335245@gmail.com

母語時制與體態標記顯性程度在二語時制習得中的作用
—難度梯度假說的實證檢驗

藤井由佳
築波大學

本研究探討顯性形態標記與跨語言形式—意義映射如何影響二語學習者對時制—體態意義的解讀。基於難度梯度假說與特徵重組假說，本研究對以日語和漢語為母語的英語學習者實施了時制—體態解讀選擇任務。結果顯示，儘管漢語缺乏語法時制標記，漢語學習者在解讀準確率上與日語學習者相當。然而，兩組學習者採用了不同的解讀策略：日語學習者主要依賴形態線索，而漢語學習者更多依賴語境資訊。此外，當母語與英語之間缺乏直接的形式—意義映射時，學習者的回答表現出更大的變異性。研究結果表明，僅憑顯性形態標記不足以充分解釋學習者的時制—體態解讀表現，學習者的表現同時受到線索可及性、跨語言形式—意義映射以及多重解讀消歧策略的共同影響。

關鍵字：難度梯度假說、跨語言形式—意義映射、時體標記的明示性、母語為日語和漢語的英語學習者、時制—體態解讀

**FROM EYE-TRACKING TO INPUT DESIGN:
ENHANCING VERB PROCESSING IN L3 SPANISH FOR
CHINESE LEARNERS**

Hui-Chuan Lu

National Cheng Kung University (Taiwan)

An Chung Cheng

University of Toledo (USA)

Zi-Xu Liu

National Cheng Kung University (Taiwan)

ABSTRACT

This study examines how Chinese-speaking learners of L3 Spanish process verb morphology during reading and how eye-tracking evidence can inform input design. Drawing on three studies from the Taiwan Eye-Movement Corpus of Spanish (TECS), the paper synthesizes findings on present/preterite tense, morphological future tense, and the *ser/estar* + *adjective* contrast. Across studies, learners frequently relied on lexical or contextual cues, such as temporal adverbs and adjectives, while under-attending to verb forms that carried critical temporal or semantic meaning. Longer fixation on verb regions was generally associated with more accurate comprehension, suggesting that attention to form supports form-meaning mapping. Corpus-level prediction and clustering analyses further identified distinct reading profiles, including Fast, Careful, and Inconsistent Readers. The findings support a cautious interpretation of noticing and input processing: eye movements indicate attentional allocation, not awareness or acquisition in and of themselves. Pedagogically, the study argues for structured input, visual enhancement, and reading-strategy support that make Spanish verb morphology visible, meaningful, and necessary for comprehension.

Keywords: eye-tracking, L3 Spanish acquisition, Spanish verb processing, input processing, structured input, reading behavior profiles

1. INTRODUCTION

In Taiwan's foreign language education context, Spanish has emerged as a popular third language (L3) option for students whose L1 is Mandarin Chinese and who have learned English as a second language (L2). Yet learning Spanish verb morphology remains a persistent challenge for these learners. Spanish is a richly inflected Romance language, encoding tense, aspect, mood, person, and number through verb endings, a feature absent in Mandarin Chinese, a morphosyntactically isolating language. Even learners with prior exposure to English, which marks tense through auxiliaries and limited suffixes, often struggle with Spanish's richer inflectional system, which requires attention to verb suffixes that convey subtle differences in meaning. Therefore, ensuring that L3 Spanish learners notice and process verb inflections is essential for accurate comprehension and acquisition.

Prior research and classroom experience suggest that Chinese speakers often struggle with Spanish verb forms (Lu et al. 2015). They may default to L1- or L2-based strategies, such as relying on time adverbs (e.g., "yesterday" or "tomorrow") or pragmatic context to infer temporal meaning, instead of attending to verb morphology. In Chinese, temporal information is typically conveyed through time phrases, aspect particles, or contextual inference rather than verb inflection (Lin 2006). Consequently, learners may initially perceive Spanish verb endings as redundant or peripheral. A student might understand *Mañana viajaré* ("Tomorrow I will travel") based on *mañana*, while overlooking the future suffix *-ré*. In the absence of such cues (e.g., *Viajaré al extranjero*, 'I will travel abroad'), the student might misinterpret the sentence's timing or default to the present tense. Such tendencies often lead to misunderstandings when adverbial cues are unavailable or when verb endings are misprocessed. These issues are compounded in L3 acquisition, where transfer effects from both L1 and L2 may influence learners' processing strategies. In particular, learners may overgeneralize from languages that express tense lexically (e.g., Chinese) or are less

morphologically complex (e.g., English), thereby reducing attention to Spanish inflections.

To investigate how L3 Chinese-speaking learners interpret Spanish verb forms during reading, we analyzed data from three previously conducted eye-tracking studies, each focusing on a different grammatical structure (present/past tense, future tense, and *ser/estar*). These experiments formed the empirical foundation for the Taiwan Eye-Movement Corpus of Spanish (TECS), which also included a recent bidirectional modeling component examining reading behavior profiles.

Eye-tracking provides a powerful means of observing L3 learners' real-time reading behavior. It allows researchers to determine which sentence elements draw visual attention and which are glossed over (Roberts and Siyanova-Chanturia 2013). Previous studies have shown that skilled readers adjust their gaze duration based on lexical and grammatical difficulty (Rayner 2009): information-rich or ambiguous words receive longer fixations, whereas predictable or low-value elements may be skipped. Thus, if Chinese-speaking learners tend to skip Spanish verb endings, especially in the absence of adverbial cues, this may reflect a processing bias that affects comprehension.

However, eye movement data must be interpreted with caution. While gaze fixations and regressions reflect where cognitive resources are allocated, they do not confirm conscious awareness or deliberate noticing (Godfroid 2020; Reichle et al. 2003). Eye-tracking captures attention deployment—an essential precondition for noticing—but should not be assumed to equate to awareness unless corroborated by behavioral outcomes or learner reports.

This paper connects eye-tracking evidence to instructional design by asking: What do learners attend to? What do they miss? And how can we shape input so that critical features (like tense markers or verb contrasts) are more likely to be noticed and processed? The next section reviews key theoretical perspectives on noticing, input processing, crosslinguistic influence, and eye-tracking in second language acquisition (SLA).

2. LITERATURE REVIEW

Understanding how learners acquire verb morphology in a third language (L3) requires integrating insights from cognitive models of attention, psycholinguistic accounts of input processing, and crosslinguistic transfer frameworks. Adult language learners do not passively perceive input; rather, their existing cognitive architecture and prior linguistic systems actively filter incoming textual data. This section highlights three key strands of research relevant to our study: the role of noticing and input processing, the theoretical framing of crosslinguistic influence, and the methodological contributions of eye-tracking in SLA.

2.1 Noticing and Input Processing

Research on second language acquisition has long shown that learners do not acquire all forms they encounter with equal ease. Some features are salient and meaning-rich; others remain largely invisible unless attention is directed at them. Schmidt's Noticing Hypothesis (1990, 2001) offers a useful starting point: input does not become intake simply because it is present. Learners must notice a linguistic feature for it to become available for further processing and learning. For Spanish, this matters because verb endings encode tense, aspect, mood, person, and number, but these meanings are only available when learners register the inflection as part of sentence interpretation. This is especially relevant when the crucial information is packed into short, low-salience endings that can easily be overshadowed by more obvious lexical material.

Later work refined the meaning of "noticing." Tomlin and Villa (1994) distinguish among attention, detection, and awareness, arguing that learners may register a form without being able to report it explicitly. This distinction matters here because it prevents a simple equation between "looking at" a form and "learning" it. In reading, learners may briefly fixate on a verb ending, skip it, or return to it after a comprehension problem. Each behavior reveals something about processing, but none alone proves that the learner has understood the grammatical value of the form. In this study, noticing is therefore treated as a form-meaning process rather than a purely visual event. The question is not only whether learners visually attend to the form, but whether learners use that form to build the sentence's meaning.

VanPatten's Input Processing theory (2004, 2025) helps explain why learners often fail to attend to morphology during comprehension. Its central claim is that learners process input primarily for meaning rather than for grammatical analysis. Under time pressure, they allocate limited attention first to elements that seem most useful for building a quick interpretation. Several familiar principles follow. The Primacy of Content Words predicts that learners attend to lexical items before grammatical endings. The Lexical Preference Principle predicts that when a lexical cue and a grammatical cue express similar information, learners rely on the lexical cue. The Preference for Non-Redundancy predicts that learners often ignore morphology if another part of the sentence has already made the meaning clear.

These principles help explain a recurring difficulty for Chinese-speaking learners of Spanish. Mandarin does not use verb inflection to express tense as Spanish does, and temporal reference is often conveyed through adverbs, particles, or discourse context (Lin 2006). A learner reading "*Ayer comí pizza*" / "Yesterday I ate pizza" may therefore interpret past time from *ayer* and never fully process the suffix *-í*. From an input-processing perspective, the problem is not simply a lack of exposure. It is a mismatch between what Spanish requires for accurate interpretation and what the learner's processing habits treat as worth attending to. This literature frames the central issue of the present study as a problem of form-meaning allocation during reading, rather than merely a problem of grammatical knowledge.

2.2 Eye-Tracking and Visual Attention

Eye-tracking allows researchers to observe how attention is distributed during reading without interrupting comprehension. In applied linguistics, it has become a useful method for examining how learners allocate attention to words, inflections, and sentence regions in real time. For this paper, the key point is simple: when readers spend very little time on a form, skip it, or return to it only after difficulty, those patterns provide evidence about how the form is being treated during comprehension (Rayner 2009; Roberts and Siyanova-Chanturia 2013).

The four gaze measures used in this study can be defined briefly. First-pass fixation duration captures the initial time spent on a target region and therefore indexes early attention when the form is first encountered. Total dwell time sums all fixations on that region and reflects later integration, reinspection, or continued difficulty. Skip rate shows how often a region is not fixated during first-pass reading; when the skipped region is a tense-bearing verb ending, it suggests that the form was treated as low-value or redundant for immediate interpretation. Regressions are backward eye movements to earlier text and usually appear when the reader needs to repair or revise an initial interpretation.

These measures are informative, but they must be interpreted cautiously. Eye movements show where attention is allocated; they do not by themselves prove conscious awareness or durable learning. Godfroid (2020) and Reichle, Rayner, and Pollatsek (2003) caution against treating eye movements as direct evidence of awareness. A longer fixation may reflect difficulty or uncertainty, and a skipped form may be predictable rather than ignored. Therefore, eye-tracking should be considered alongside comprehension outcomes and learner variables.

Grammar-focused eye-tracking research supports this use of the method. Winke (2013), in a study of textual enhancement with eye-movement data, examined whether visually enhanced forms changed attention to grammar and learning outcomes. Indrarathne and Kormos (2017) found that learners who were explicitly directed to attend to a target syntactic construction showed stronger attentional processing and larger gains than learners in less explicit conditions. Together, these studies justify using eye-tracking here as a window on selective attention to Spanish verb morphology: not as proof of acquisition, but as evidence about whether learners give grammatical forms a real chance to contribute to comprehension.

2.3 Crosslinguistic Influence and L3 Verb Processing

Because the learners in this study are multilingual, their reading strategies are shaped by at least two prior language systems. L3 acquisition research has shown that previously acquired languages can influence a new language in selective and sometimes competing ways

(Cenoz 2003; De Angelis 2007). For Chinese-speaking learners of Spanish in Taiwan, both Mandarin Chinese and English matter, but for different reasons. Mandarin contributes entrenched habits for interpreting time and event structure without verbal inflection. English contributes familiarity with alphabetic print, subject-verb-object order, and some tense marking, but it does not prepare learners for the density of Spanish conjugation.

L3 models help clarify this situation. Rothman's Typological Primacy Model proposes that learners initially draw on the prior language they perceive as most closely typologically related to the target. The Linguistic Proximity Model argues instead that transfer may occur on a property-by-property basis rather than wholesale (Rothman 2011; Westergaard et al. 2017). For the learners in this study, English may appear closer to Spanish in the writing system and broad sentence structure. That may help with word recognition and segmentation. Yet when the relevant property is verbal morphology, neither prior language offers a close match. Mandarin lacks inflectional tense marking, and English uses a comparatively light system that often depends on auxiliaries or a small set of endings. Learners may thus enter Spanish reading with some transferable resources, but not with a routine that treats verb endings as central cues to meaning.

This mismatch is best understood as both a processing and a representational issue. Learners may know that Spanish verbs change form for tense or person, but under reading pressure, they can still fall back on older routines: scan for time adverbs, identify key nouns, extract gist from lexical items, and move on. That strategy works in Mandarin and often well enough in English. In Spanish, however, it can lead to shallow interpretations or misunderstandings when inflection carries information that the lexical context does not supply.

The same logic applies to structures such as *ser* and *estar*. Neither Mandarin nor English provides a direct equivalent to the Spanish distinction, so learners may default to the broad meaning "be" and rely on the adjective alone for interpretation. Spanish verb morphology is therefore not just another grammar point. It is a location where prior-language habits, perceptual salience, and limited attentional resources converge. This makes Chinese-speaking learners of L3 Spanish an

especially informative population for examining how crosslinguistic influence shapes what is noticed, skipped, and understood during reading.

2.4 Research Gaps and Rationale for the TECS Project

While eye-tracking has gained traction in second-language acquisition (SLA) research, studies examining how Chinese-speaking learners process Spanish verb morphology remain extremely limited. Much of the prior work has either emphasized learner production (e.g., corpus-based analyses of the past tense, Lu, Cheng, and Hung 2015) or focused on Indo-European learner populations with linguistic backgrounds different from those of Mandarin Chinese speakers. These approaches offer valuable insights into output patterns but fall short in revealing real-time attentional dynamics during input processing.

An important step toward addressing this gap is the work of Lu and Cheng (2025), who used eye-tracking to examine how Chinese-speaking learners process Spanish future-tense verbs. Their study showed that learners relied on temporal adverbs (e.g., *mañana*, *el próximo año*) to support future-time interpretation, while their sensitivity to future-tense morphology was still developing. This suggests that learners may not notice verb morphology unless prompted by other contextual cues, highlighting the need for targeted pedagogical interventions.

To build on and extend this line of inquiry, the Taiwan Eye-Movement Corpus of Spanish (TECS) was developed as a multi-year project to systematically examine how Mandarin-speaking learners of Spanish process verb-related grammatical structures. The current TECS project comprises three controlled eye-tracking experiments, each targeting a distinct linguistic feature (e.g., tense contrasts, unfamiliar morphology, and copula usage) alongside corpus-level modeling of learner reading behaviors. This design allows for both focused analysis of morphosyntactic attention and broader predictions about learner proficiency and reading profiles. By analyzing learners' gaze patterns and comprehension outcomes, TECS provides a data-driven foundation for designing instructional strategies that better align with real-time processing constraints.

The current study draws upon these three experiments and the TECS bidirectional model to synthesize findings and propose pedagogical implications. It addresses the empirical gap in L3 input processing research among Chinese-speaking learners and contributes to the growing effort to make eye-tracking insights actionable for language teaching.

3. METHODOLOGY

This paper synthesizes data from three independently conducted eye-tracking studies carried out from 2021 to 2023 as part of the Taiwan Eye-Movement Corpus of Spanish (TECS). Each study examined how Mandarin-speaking learners of Spanish processed a different verb-related structure during reading: present versus preterite tense, morphological future tense, and the contrast between *ser* and *estar* with adjectives. Although the studies were conducted in different academic years and with different cohorts, they shared a common methodological orientation: learners read Spanish sentences or passages while their eye movements were recorded, and they then completed comprehension or interpretation questions. This cumulative design allows the present paper to compare patterns of attention, comprehension, and form-meaning mapping across multiple grammatical domains.

3.1 Research Questions

The study addresses the following research questions:

- (1) How do Chinese-speaking L3 learners of Spanish allocate attention to verb-related forms during reading?
- (2) How does attention to Spanish verb morphology relate to comprehension accuracy across tense and copular-verb structures?
- (3) How can eye-tracking and corpus-based findings inform the design of input activities that enhance learners' attention to Spanish verb morphology?

3.2 Participants and Shared Procedures

Across the three studies, participants were undergraduate students in Taiwan enrolled in Spanish language courses. All participants were native speakers of Mandarin Chinese, had learned English as an L2, and were studying Spanish as an additional language. Their Spanish proficiency ranged approximately from beginner to intermediate levels, depending on the year and study. Each study used a standardized proficiency measure, including the Avant STAMP 4S assessment, to document learners' Spanish proficiency and, where relevant, to describe or classify learner groups for later analysis.

The eye-tracking experiments were conducted in controlled laboratory settings. Participants sat in front of a computer screen while a desktop-mounted infrared eye-tracker recorded their eye movements. The eye-tracker was calibrated individually for each participant before the experiment began. During the tasks, participants read Spanish stimuli silently and then answered comprehension or interpretation questions. The target regions, or Areas of Interest (AOIs), were defined according to the grammatical focus of each study: verb endings in the tense studies, future-tense morphemes in the future-tense study, and copular verbs and adjectives in the *ser/estar* study. The primary eye-movement measure used across studies was total fixation duration, supplemented where relevant by fixation patterns on surrounding lexical cues such as temporal adverbs or adjectives.

Because the purpose of the present article is synthetic rather than experimental replication, the following subsections provide a concise overview of each study's design, participants, materials, and focus.

3.3 Overview of the Three Eye-Tracking Studies

Study 1: Present versus Preterite Tense

Study 1 examined how learners processed Spanish present-tense and preterite verb forms during reading. The final analysis included 34 participants. The stimuli included sentences with verbs in either the present or preterite tense. Some sentences contained temporal adverbs that supported tense interpretation, while others did not. This design allowed us to compare learners' attention to verb endings when temporal

information was available lexically and when the verb ending was the primary cue.

After reading each sentence, participants answered a comprehension or grammaticality judgment question. The critical AOIs were the tense-bearing endings of the target verbs and, when present, the temporal adverbs. This study provided evidence about whether learners interpreted temporal reference through Spanish verb morphology or relied more heavily on lexical time cues.

Study 2: Morphological Future Tense

Study 2 focused on the Spanish morphological future tense, a form that many participants had not yet formally learned at the time of testing. The final analysis included 37 participants. Participants read short Spanish texts containing future-tense verb forms, such as forms ending in *-rá* or related future-tense endings. Some texts also contained temporal adverbs or expressions signaling futurity, while others required learners to infer futurity from the verb morphology itself.

After each text, participants answered questions assessing their understanding of future reference and their selection of appropriate future-tense forms. The critical AOIs were the future-tense verb endings and the temporal adverbs in texts with lexical future-time cues. This study enabled examination of how learners respond to a meaningful but unfamiliar morphological form, and whether lexical temporal cues support or compete with attention to future-tense morphology.

Study 3: *Ser/Estar* + Adjective

Study 3 examined how learners processed the semantic contrast between *ser* and *estar* in adjective constructions. The final analysis included 39 learners of Spanish, along with a small native-speaker comparison group used descriptively. Participants read sentences containing either *ser* or *estar* followed by an adjective. Each sentence was followed by contextual interpretation options, requiring learners to select the context that best matched the meaning of the Spanish sentence.

The critical AOIs included the subject, the copular verb, and the adjective. Because *ser* and *estar* differ in length, fixation duration on *ser* forms was adjusted to allow a fairer comparison with *estar*. This study allowed us to examine whether learners attended to the copular verb as a

meaningful cue or relied primarily on the adjective to interpret the sentence.

3.4 TECS Corpus-Level Analysis and Prediction Models

In addition to synthesizing the three experimental studies, the present paper draws on corpus-level analysis from TECS. The goal of this analysis was to determine whether learners' eye-movement behavior could be used not only to interpret individual experimental results but also to identify broader reading behavior patterns across tasks and learner groups.

The TECS analysis included two complementary prediction models developed in the project: a unidirectional prediction model linking eye-movement indices to comprehension accuracy (Lu et al. 2024) and a bidirectional model relating gaze behavior and proficiency (Lu, Cheng, and Ho 2025). The first was a unidirectional model, in which eye-movement indices were used to predict learners' comprehension accuracy. This model addressed whether gaze behavior could serve as an indicator of successful or unsuccessful sentence interpretation. The second was a bidirectional model, which examined the relationship between learners' eye-movement behavior and Spanish proficiency in both directions: gaze behavior was used to predict proficiency, and proficiency was used to predict expected gaze patterns. Together, these models allowed us to examine whether reading behavior, comprehension, and proficiency were systematically related across the TECS dataset.

The unidirectional model examined whether eye-movement indices could predict learners' comprehension accuracy.

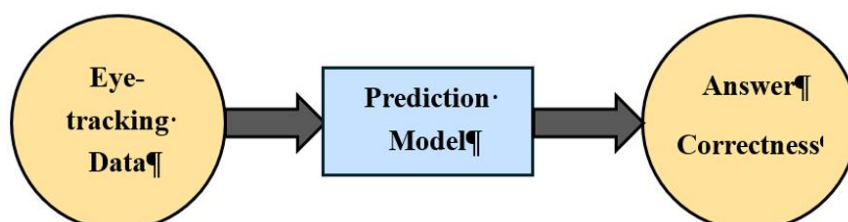


Figure 1. Unidirectional prediction model linking eye-tracking measures to comprehension accuracy.

The bidirectional model examined the relationship between gaze behavior and proficiency in both directions.

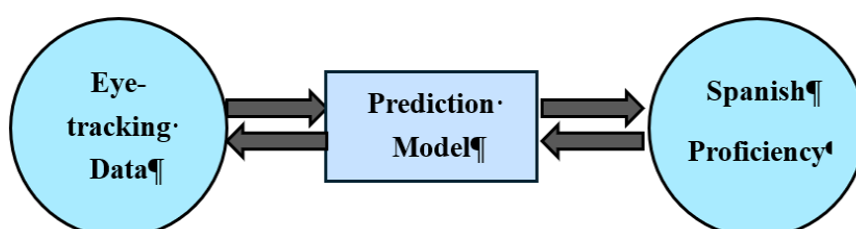


Figure 2. Bidirectional prediction model linking eye-movement patterns and proficiency level.

3.5 Clustering Analysis

The TECS corpus-level analysis also included a K-means clustering of learner reading behaviors. Based on gaze features such as fixation duration, verb fixation ratio, and attention distribution across verbs and lexical cues, learners were grouped into reading profiles. To explore the stability and interpretability of these profiles, clustering analyses were conducted on both the combined dataset and datasets separated by comprehension outcomes (correct vs. incorrect responses). Cluster quality was evaluated using silhouette coefficients. These profiles are discussed in Section 4.3, where they are interpreted in relation to comprehension outcomes and the design of pedagogical input.

4. FINDINGS AND DISCUSSION

The eye-tracking data across the three studies indicate that Spanish verb morphology is a major bottleneck in L3 reading for Chinese-speaking learners. Although the studies differed in grammatical focus (present/preterite tense, future tense, and *ser/estar* with adjectives), a consistent pattern emerged: learners often relied on lexical or contextual cues while under-attending to verb forms. This tendency affected comprehension most clearly when sentence interpretation depended on morphology rather than on temporal adverbs, adjectives, or general context.

This section presents the findings in four parts. Section 4.1 summarizes the major findings from the three eye-tracking studies. Section 4.2 discusses the shared pattern of lexical-cue reliance and its consequences for form-meaning mapping. Section 4.3 presents the TECS corpus-level prediction models and reader profiles. Section 4.4 synthesizes the findings regarding noticing, input processing, and crosslinguistic influence.

4.1 Verb Form Attention and Comprehension across the Three Eye-Tracking Studies

Study 1 examined how learners processed present and preterite verb forms, both with and without temporal adverbs. In sentences without temporal adverbs, correct tense interpretation was strongly associated with longer fixation on verb endings. For present-tense verbs, learners who interpreted the sentence correctly spent more time on the ending of the verb than those who answered incorrectly. The same pattern was even stronger for preterite verbs. Among correctly interpreted items, preterite endings also required longer fixation than present-tense endings. These results suggest that learners who devoted more attention to the verb ending were more likely to accurately identify the sentence's temporal reference.

The results from Study 1 also showed that temporal adverbs shaped how learners allocated attention. When a temporal adverb appeared before the verb, it often provided an early lexical cue that reduced the need to

attend to the verb ending. In past-tense congruent sentences, for example, learners fixated longer on the verb ending when the verb appeared before the adverb than when the adverb appeared first. This pattern suggests that when a salient time expression is encountered early, learners may use it to establish the temporal frame and then devote less attention to the morphological cue. In other words, temporal adverbs can support comprehension, but they can also reduce learners' attention to verb morphology.

Study 2 extended this issue to the Spanish morphological future tense, which many participants had not formally learned at the time of testing. The findings showed that learners' sensitivity to the future-tense ending was still developing. In texts containing temporal adverbs, Spanish proficiency significantly predicted accuracy in future-tense selection, while fixation on future-tense endings approached significance. For near-intermediate learners, future-tense selection accuracy was higher after reading texts with temporal adverbs than after reading texts without them. In addition, several comparisons showed that learners fixated longer on future-tense verb endings in the later texts that contained temporal adverbs. These findings suggest that temporal adverbs helped learners interpret future reference, while repeated exposure to future-tense endings may have supported emerging sensitivity to the unfamiliar morphology.

At the same time, Study 2 should not be interpreted as showing that learners completely ignored future-tense morphology. Rather, the data suggest a more nuanced pattern: learners relied on temporal adverbs when they were available, but some learners, especially those with higher proficiency, began to attend to the future-tense ending as a meaningful cue. This is important pedagogically because it shows that lexical support can serve as a bridge to comprehension, but instruction must eventually guide learners beyond lexical reliance toward morphological interpretation.

Study 3 examined the distinction between *ser* and *estar* in adjective constructions. Unlike Studies 1 and 2, this study focused on verb choice rather than tense suffixes, but the processing problem was similar. Learners often relied on the adjective and treated both verbs as broadly equivalent to "be," even though *ser* and *estar* guide different interpretations of the adjective. The results showed significantly higher accuracy for *ser* + *adjective* sentences than for *estar* + *adjective* sentences.

In correctly interpreted responses, learners also fixated longer on *estar* than on weighted *ser* forms, suggesting that *estar* required greater processing effort even when learners selected the correct context.

Study 3 also showed that correct interpretation was associated with greater attention to the copular verb. In both *ser* and *estar* structures, learners who answered correctly spent more time fixating on the verb than those who answered incorrectly. This finding is particularly important because it shows that the copular verb was not a semantically empty connector in Spanish input. For these learners, however, prior experience with Chinese and English may have encouraged them to treat copular forms as low-value elements and to focus instead on the adjective. The eye-tracking data, therefore, reveal a specific form-meaning mapping problem: learners must learn that the choice between *ser* and *estar* affects how the adjective should be interpreted.

Taken together, the three studies show that learners did not ignore all verbs in the same way. Rather, they under-attended to the specific part of the verb that carried the grammatical or semantic contrast needed for interpretation: tense endings in Studies 1 and 2, and copular choice in Study 3. This distinction is important because it shows that the difficulty is not simply a matter of vocabulary knowledge. Learners may understand the verb stem, the adjective, or the general sentence meaning yet still fail to process the grammatical cue that determines the intended interpretation.

To clarify the cross-study pattern, Table 1 summarizes the three eye-tracking studies in parallel. Rather than presenting each experiment as an isolated case, the table highlights the target structure, the participant group, the key quantitative finding, and the pedagogical implications for each study. This comparison shows that, despite differences in grammatical focus, all three studies point to the same processing issue: learners' comprehension improves when they allocate attention to verb forms rather than relying exclusively on lexical or contextual cues.

Table 1. Summary of Eye-Tracking Studies on Spanish Verb Processing

Study	Target structure	Participants	Key quantitative finding	Pedagogical implication
Study 1	Present vs. preterite	n = 34	Correct tense interpretation was associated with longer fixation on verb endings; present: 0.81 s vs. 0.25 s, $p < .001$; preterite: 1.45 s vs. 0.12 s, $p < .001$.	Reduce overreliance on temporal adverbs; design input tasks that require attention to tense endings.
Study 2	Future tense	n = 37	For near-intermediate learners, future-tense accuracy improved after texts with temporal adverbs: 5.58 vs. 4.52, $p = .026$.	Use temporal adverbs as scaffolds, then gradually remove them so learners attend to future endings.
Study 3	<i>Ser/estar</i> + adjective	n = 39	Accuracy was higher for <i>ser</i> than <i>estar</i> : 0.82 vs. 0.56, $p < .001$. In correctly interpreted responses, fixation was longer on <i>estar</i> than on weighted <i>ser</i> forms: 1.24 s vs. 0.96 s, $p < .001$.	Use contextual interpretation tasks that make copular choice necessary for meaning.

4.2 Lexical-Cue Reliance and Form-Meaning Mapping

A central pattern across Studies 1 and 2 is that temporal adverbs helped learners understand sentence time, but also encouraged reliance on lexical cues. In Study 1, learners were more likely to interpret tense correctly when temporal adverbs were available, but the same adverbs could reduce attention to verb endings. In Study 2, temporal adverbs supported future-time interpretation, especially for near-intermediate learners, but learners' sensitivity to future-tense morphology remained incomplete.

This pattern reflects a central dilemma in L3 Spanish teaching. Temporal adverbs make input more comprehensible, especially for learners whose L1 expresses time through lexical or contextual means. However, if learners consistently rely on adverbs to determine time reference, they may not develop stable form-meaning mappings between Spanish verb endings and temporal interpretation. Apparent comprehension may therefore mask incomplete processing. A learner may correctly understand that an event is in the past because of *ayer*, or in the future because of *mañana*, without processing the corresponding verb ending.

Study 3 shows that the same issue extends beyond tense. In *ser/estar* + *adjective* constructions, the adjective is often the most semantically transparent element. Learners may therefore construct a plausible interpretation from the adjective alone and overlook the copular verb. However, in Spanish, the copula contributes to how the adjective is interpreted. When learners do not attend to *ser* or *estar*, they may understand the general descriptive meaning but miss the intended contrast between a general property and a situated condition.

These findings support the claim that Spanish verb morphology functions as a comprehension bottleneck for Chinese-speaking L3 learners. The bottleneck is not only formal but functional. Learners must learn that Spanish verb forms are not decorative, redundant, or merely grammatical labels. They carry meaning that may not be recoverable from lexical context alone.

At the same time, the findings should be interpreted with caution. Eye-tracking data show where attention is allocated, but they do not directly

prove conscious noticing or long-term learning. Longer fixation may reflect difficulty, uncertainty, or successful integration; shorter fixation may reflect efficiency, predictability, or neglect. For this reason, the present study interprets gaze behavior together with comprehension outcomes. When reduced attention to verb forms coincides with lower comprehension, the evidence becomes stronger that form neglect has consequences for interpretation.

4.3 TECS Prediction Models and Reading Behavior Profiles

The TECS corpus-level analysis extends the findings from the three individual studies by examining broader patterns in learner reading behavior. Rather than asking only whether learners processed a particular verb form in a single experiment, the TECS analysis asks whether learners show recurring reading profiles across tasks and grammatical structures.

The TECS project developed two complementary prediction models. The first was a unidirectional model, in which eye-movement indices were used to predict learners' comprehension accuracy. This model addressed whether gaze behavior could serve as an indicator of successful or unsuccessful sentence interpretation. The second was a bidirectional model that examined the relationship between learners' eye-movement behavior and Spanish proficiency in both directions: gaze behavior predicted proficiency, and proficiency predicted expected gaze patterns. This bidirectional logic is useful because it treats reading behavior and proficiency as mutually informative rather than as separate outcomes.

Table 2 summarizes how the TECS corpus-level analyses extend the findings from the three eye-tracking studies. The models do not replace the experimental findings; rather, they provide a broader view of how gaze behavior, comprehension, and proficiency are related across the corpus. In particular, the bidirectional model and clustering analysis help identify recurring reading profiles that have direct implications for input design.

Table 2. TECS Corpus-Level Prediction and Reading Profiles

Analysis	Prediction direction / method	Key result	Pedagogical implication
Unidirectional model	Eye-movement indices → comprehension accuracy	Gaze behavior was used to predict whether learners interpreted sentences correctly.	Eye-tracking measures can help identify processing breakdowns during reading.
Bidirectional model	Eye movement ↔ Spanish proficiency	Gaze behavior predicted proficiency, and proficiency predicted expected gaze patterns.	Reading behavior and proficiency can be treated as mutually informative.
Clustering analysis	K-means clustering of gaze features	Three profiles emerged: Fast Readers, Careful Readers, and Inconsistent Readers; silhouette coefficients ranged from 0.36 to 0.39.	Input design can be differentiated by learner profile: slowing down Fast Readers, building fluency for Careful Readers, and stabilizing strategies for Inconsistent Readers.

The clustering analysis identified three broad reader profiles: Fast Readers, Careful Readers, and Inconsistent Readers. The reader profiles were derived from features such as verb fixation ratio, total verb fixation duration, and fixation distribution across adverbs or adjectives. Across yearly datasets, the clustering models showed average silhouette

coefficients above 0.35, suggesting that the model captured reasonably distinct behavioral patterns despite the natural variability of eye-movement data.

To further evaluate the interpretability of the profiles, clustering solutions were compared across combined datasets and datasets separated by comprehension outcomes (correct vs. incorrect responses).

The profiles help clarify learner differences that are not fully explained by proficiency alone. Fast Readers moved quickly through sentences and often showed short total fixation times on critical regions. They could sometimes grasp the general meaning of a sentence, especially when lexical cues were available, but they were more likely to miss grammatical details. Careful Readers read more slowly and tended to allocate more sustained attention to relevant sentence regions, including verbs. Their reading behavior was generally associated with more successful interpretation. Inconsistent Readers showed fewer stable patterns; their attention to verbs and contextual cues varied across items, making their comprehension less predictable.

One important result from the TECS clustering analysis is that speed alone should not be taken as evidence of efficient comprehension. Some Fast Readers allocated a relatively high proportion of attention to verbs, but when comprehension was incorrect, their absolute fixation time on verbs was lower than that of Fast Readers who answered correctly. This suggests that both proportion and duration matter. Learners may know where to look, but if they do not spend enough time processing the form, comprehension may still fail.

The TECS model, therefore, contributes beyond individual experiments. It shows that verb processing difficulties are part of broader reading behavior patterns. It also provides a diagnostic bridge between research and teaching. If gaze behavior can help identify learners who consistently move too quickly past critical verb cues, teachers can design activities that promote slower and more deliberate attention to form. If learners already attend carefully but remain inaccurate, instruction may need to focus more on semantic interpretation or explicit form-meaning mapping. Thus, the value of the TECS model lies not only in prediction but also in its ability to guide differentiated input design.

4.4 Cross-Study Synthesis and Theoretical Integration

The findings across the three eye-tracking studies and the TECS corpus-level analysis converge on a shared conclusion: Chinese-speaking L3 learners of Spanish often process sentences in ways that prioritize lexical meaning over grammatical morphology. Temporal adverbs, adjectives, and general context tend to attract attention more readily than tense endings or copular distinctions. This pattern helps explain why learners may understand the gist of a sentence but miss essential grammatical meanings for accurate interpretation.

These findings align closely with VanPatten's Input Processing theory. Learners' reliance on adverbs and content words reflects the Primacy of Content Words and the Lexical Preference Principle. When a lexical item already provides temporal information, learners may not allocate sufficient processing resources to the verb ending. Study 1 illustrates this in processing the present and preterite tenses, and Study 2 extends the pattern to the future tense. Study 3 shows that the same principle applies to copular verbs: when adjectives carry the most obvious semantic content, learners may under-process the verb that shapes how the adjective should be interpreted.

The findings also support a cautious interpretation of Schmidt's Noticing Hypothesis. Learners who visually attended to critical verb cues were more likely to interpret sentences correctly, suggesting that attention to form plays an important role in comprehension. However, the present study does not claim that fixation equals conscious noticing or acquisition. Rather, eye-tracking provides indirect evidence of attention, and comprehension responses indicate whether that attention supports interpretation. In this sense, the study offers evidence for the role of attention in form-meaning mapping while avoiding the stronger claim that gaze behavior directly proves learning.

Cross-linguistic influence further explains why these processing patterns are persistent. Mandarin Chinese does not require readers to attend to verb inflections for tense marking, and English provides only limited preparation for Spanish morphology. As a result, learners may enter L3 Spanish reading with established strategies that favor word order, adverbs, and lexical meaning over suffixes or subtle verb contrasts. These

strategies are useful in Mandarin and often sufficient in English, but they can lead to shallow processing in Spanish.

The TECS reading profiles add an individual-differences dimension to this theoretical picture. Not all learners under-attended to verb morphology in the same way. Fast Readers tended to move through input quickly and risk missing grammatical details. Careful Readers showed more sustained attention and generally stronger comprehension. Inconsistent Readers lacked stable processing routines. These profiles suggest that instruction should not only present verb morphology more clearly but also help learners develop better reading strategies.

Overall, the findings show that Spanish verb morphology is a point where attention, input processing, crosslinguistic influence, and learner variability intersect. This is why pedagogical intervention is necessary. Learners need input that makes verb morphology visible, meaningful, and necessary for comprehension. The next section develops this implication by proposing structured input activities, visual enhancements, and strategy-based support that respond directly to the processing patterns identified in the TECS project.

5. PEDAGOGICAL IMPLICATIONS

One of the central challenges in applying psycholinguistic research, such as eye-tracking data from the TECS corpus, to language pedagogy lies in translating empirical findings into teachable actions. It is crucial to understand how L2 learners process input and how instruction can redirect their attention to the grammatical forms they tend to overlook. The findings of the TECS project demonstrate that difficulties in Spanish verb processing are not simply a matter of grammatical knowledge. Rather, they reflect how learners allocate attention during reading. Across the three eye-tracking studies, many learners relied heavily on lexical cues while under-attending to verb morphology. This pattern was especially evident when temporal adverbs or semantically rich adjectives were available, reducing the likelihood that learners would process the grammatical information encoded in verb forms.

These findings suggest that effective instruction should not focus only on explaining grammatical rules. It should also help learners develop processing routines that direct attention to form-meaning relationships during comprehension. The pedagogical recommendations below are grounded in the TECS findings and informed by Schmidt's Noticing Hypothesis and VanPatten's Input Processing framework. Together, these perspectives suggest that learners must attend to grammatical forms before those forms can reliably support comprehension and acquisition.

5.1 Structured Input: Making Verb Forms Necessary for Meaning

Findings from Studies 1 and 2 showed that learners frequently relied on temporal adverbs such as *ayer*, *mañana*, or *el próximo año* while overlooking verb morphology. When adverbial cues were absent, comprehension accuracy declined, indicating that many learners had not yet developed stable form-meaning mappings between verb endings and temporal interpretation.

One instructional response is to use structured input activities derived from Processing Instruction (VanPatten and Cadierno 1993; VanPatten 1996; Lee and VanPatten 2003; Farley 2005). Structured input tasks are designed so that learners must interpret meaning by attending to the target grammatical form. At this stage, students are not asked to produce the form. Instead, they are asked to process the form to understand the sentence correctly. This distinction is important because the goal is to reshape how learners process input before asking them to use the form in speaking or writing.

For example, to help learners distinguish present and preterite forms, the instructor can remove explicit time adverbs and ask learners to interpret the time reference from the verb form alone:

Past vs. Present Tense (Reading Task): Read each sentence and determine whether it describes something Juana did in the past. Circle Sí (yes) if it happened in the past, or No if not.

- | | |
|--|----------------|
| (1) <i>Juana caminó a la escuela.</i> | <i>Sí / No</i> |
| (2) <i>Juana baila en la escuela.</i> | <i>Sí / No</i> |
| (3) <i>Juana trabajó hasta las diez.</i> | <i>Sí / No</i> |

In this task, learners cannot rely on a time adverb such as *ayer*. They must attend to the contrast between forms such as *caminó* and *baila*. This directly responds to the processing pattern observed in Study 1, where learners tended to use temporal adverbs as shortcuts and often skipped or under-processed the verb ending.

A similar structured input activity can be used for the future tense, which was the focus of Study 2. Because many learners had not yet formally learned the morphological future, they tended to rely on adverbs such as *mañana* or *el próximo año*. An input activity can gradually remove those supports:

Present vs. Future Tense (Yes/No Guessing Game): Your teacher is telling you about possible activities next week. Decide whether each sentence refers to a future event.

- | | |
|--|----------------|
| (1) <i>Ella se levantará a las seis de la mañana.</i> | <i>Sí / No</i> |
| (2) <i>Ella cocina el desayuno a las seis y media.</i> | <i>Sí / No</i> |
| (3) <i>Ella mirará los programas de la televisión.</i> | <i>Sí / No</i> |

These activities require learners to process verb endings as meaningful cues. Over time, such input practice can help learners move from relying mainly on lexical cues to using Spanish morphology in sentence interpretation.

5.2 Reducing Overreliance on Lexical Cues

A central finding from Studies 1 and 2 is that temporal adverbs can enhance comprehension but also reduce attention to verb morphology. This creates a pedagogical dilemma. On the one hand, adverbs such as *ayer* and *mañana* make input more comprehensible. On the other hand, if learners always use those adverbs to determine time reference, they may never learn to process tense morphology efficiently.

Instruction can address this problem by gradually reducing lexical support. In early stages, teachers may include time adverbs to establish meaning. Once learners understand the temporal contrast, the adverbs can be removed, leaving learners to rely on verb forms. For example, a

sequence of activities might begin with *Ayer Juana caminó al parque* and later move to *Juana caminó al parque*. This gradual removal of lexical cues helps learners shift attention from the more obvious time expression to the less salient verb ending.

Distinguishing Habitual vs. Past (Listening Task): Listen to each sentence the teacher reads out. Decide if it describes something Juana does every day or something she did yesterday. Write a check next to the correct time adverb for each sentence.

- | | |
|---|--------------------------------------|
| (1) <i>Juana caminó al parque.</i> | <i>todos los días/every day</i> ____ |
| | <i>ayer/yesterday</i> ____ |
| (2) <i>Juana baila en el auditorio.</i> | <i>todos los días/every day</i> ____ |
| | <i>ayer/yesterday</i> ____ |
| (3) <i>Juana trabajó hasta las cinco.</i> | <i>todos los días/every day</i> ____ |
| | <i>ayer/yesterday</i> ____ |

In all the above activities, the only way learners can get the correct answers is by paying attention to the verb conjugations (e.g., *-a* vs. *-ó*). These tasks push learners to process form and meaning together, verifying their interpretations through feedback. For instance, begin with only third-person singular forms of regular *-ar* verbs (present vs. preterite), so learners are comparing just two endings. Once students build the correct form-meaning mapping in the third-person conjugated form, the practice can move on to other forms, such as first-person, one at a time. Instructors should also provide immediate feedback in these exercises to reinforce the correct form-meaning connections. Over time, such PI tasks recalibrate learners' processing strategies.

Different target structures may require tailored input tasks. For instance, our eye-tracking experiments also examined the copular verbs *ser* and *estar* + *adjectives*, which pose a distinct challenge: beyond tense differences, *ser* vs. *estar* carries aspectual nuance that interacts with adjective semantics. Thus, a structured input approach for *ser/estar* should incorporate richer context to highlight subtle differences in meaning.

5.3 Teaching *Ser* vs. *Estar* via Form-Meaning Mapping

TECS Study 3 focused on the distinction between *ser* and *estar* with adjectives. The findings showed that learners often relied on the adjective and treated both verbs as a general equivalent of “be.” This suggests that learners may not process the semantic contribution of the copula unless the task requires it.

For this reason, instruction should avoid presenting *ser* and *estar* only as memorized rules, such as “permanent vs. temporary.” Although useful as a starting point, that explanation is too simple. The contrast often involves how the speaker construes a property, condition, state, or situation in context. Learners, therefore, need activities that connect the verb choice with the intended interpretation. This recommendation is consistent with prior Processing Instruction research showing that structured input can support learners’ interpretation of *ser* and *estar* (Cheng 2002).

One useful activity is context matching. Learners read a sentence and choose the context that best matches its meaning:

Context Matching: For each Spanish sentence, choose the context (A or B) that best matches its meaning, whether each sentence describes a general trait or a temporary state/condition.

- (1) *El hombre está activo en la cancha.*
 - a. In this basketball game, the man’s movements are especially quick and agile; he alternates between offense and defense and is essentially the highest-scoring player on the team.
 - b. On the court, the man is very proactive and enthusiastic about every kind of sport. Whenever a game needs an extra player, you can always count on him to join.
- (2) *El pastel de chocolate negro es delicioso en esta pastelería.*
 - a. Every dark chocolate cake at this pastry shop is extremely tasty. Each one is baked by an experienced pastry chef using top-quality ingredients and decorated with fresh fruit.
 - b. I bought a dark chocolate cake that was just baked, and it tastes especially delicious right now; it has hazelnut pieces in the filling and fresh fruit on top, and its surface is perfectly smooth and enticing.

In these context-matching items, only one of the two scenarios corresponds appropriately to the given Spanish sentence (e.g., the dynamic, in-the-moment scenario matches *está activo*, whereas the general habit scenario would require *es activo*). Students must identify which copular verb is used and understand its implications to choose correctly. This kind of activity prompts them to pay attention to the verb (*ser* vs. *estar*) in relation to contextual cues, reinforcing the form-meaning linkage for copular verbs. To deepen form-meaning connections, tasks should avoid direct translation and instead focus on interpretation. The TECS findings underscore that verb semantics must be made both salient and meaningful to be processed effectively.

This type of activity aligns directly with Study 3. Learners must attend not only to the adjective but also to the copular verb. In doing so, they practice mapping the verb choice onto contextual meaning. The goal is not simply to memorize a rule, but to learn how *ser* and *estar* guide interpretation.

5.4 Using Visual Input Enhancement as a Temporary Scaffold

Visual input enhancement can also support noticing, especially at early stages (Sharwood Smith 1993; Lee and Huang 2008). Teachers may boldface, underline, or color-code target forms such as *-ó*, *-ará*, *ser*, or *estar*. For example, a reading passage might highlight preterite endings in one color and present-tense endings in another. This can help learners notice forms they might otherwise skip.

However, visual enhancement should be used carefully. Prior research suggests that textual enhancement can increase the salience of target forms, but its effects on acquisition are mixed and depend on how enhancement is integrated into meaningful input and tasks (Han, Park, and Combs 2008). Highlighting a form may draw the learner's eyes to it, but it does not guarantee that the learner understands its function. Therefore, input enhancement should be paired with interpretation tasks. After reading an enhanced passage, learners might answer questions that require them to use the highlighted form to determine when an event occurred or how a description should be interpreted.

In this sense, visual enhancement is best understood as a scaffold rather than a complete instructional solution. It can make the target form more visible, but structured input activities are needed to make the form meaningful.

5.5 Addressing Different Reading Profiles

The TECS corpus-level analysis also identified three reading profiles, including Fast Readers, Careful Readers, and Inconsistent Readers. These profiles are pedagogically relevant because not all learners need the same kind of support.

Fast Readers may understand the general meaning of a sentence but miss grammatical details. They may benefit from tasks that require them to slow down and verify verb forms before answering comprehension questions. Careful Readers may already attend to grammatical details but need support in developing fluency and confidence. Inconsistent Readers may require explicit strategy training, such as rereading the verb when the sentence meaning is unclear.

Teachers can make these strategies explicit. For example, students can be reminded: “If the time reference is unclear, check the verb ending,” or “If a sentence with *ser* or *estar* seems ambiguous, reread the verb and adjective together.” Such strategies translate eye-tracking findings into classroom routines that learners can apply independently.

5.6 From Eye-Tracking Evidence to Input Design

The main pedagogical implication of this study is that input must be designed so that learners cannot bypass the grammatical form and still arrive at the intended meaning. The TECS findings show that Chinese-speaking learners of L3 Spanish often rely on lexical cues, skip or underprocess verb morphology, and vary in reading strategy. Structured input, contextualized interpretation tasks, visual enhancement, and self-monitoring strategies all respond to these findings in different ways.

These recommendations do not imply that instruction should become mechanical or form-only. Communicative and meaning-based instruction remains essential. The point is that form and meaning must be connected

during comprehension. By designing input that guides learners' attention to verb morphology, teachers can help learners process Spanish in ways that better align with how the language encodes meaning.

In this sense, the movement from eye-tracking to input design is not simply a methodological contribution. It is a practical pathway for helping teachers identify where learners' attention breaks down and how classroom materials can be adjusted to support more accurate and meaningful L3 Spanish comprehension.

6. CONCLUSION

This study addressed the overarching research question of how eye-tracking and corpus-based findings can inform the design of input activities that enhance Chinese-speaking learners' attention to Spanish verb morphology during L3 reading. By synthesizing three eye-tracking studies from the Taiwan Eye-Movement Corpus of Spanish (TECS), together with corpus-level prediction and clustering analyses, the study examined how learners process verb-related forms across present/preterite tense, morphological future tense, and the *ser/estar* + *adjective* contrast.

Across the three studies, a consistent pattern emerged: learners often relied on lexical or contextual cues while under-attending to verb forms that carried critical temporal or semantic information. Temporal adverbs supported interpretation in the tense studies, but they also reduced the need to process verb endings. In the *ser/estar* study, learners often relied on adjectives while overlooking the semantic contribution of the copular verb. The TECS reading-profile analysis further showed that learners differed in how they allocated attention during reading, with Fast, Careful, and Inconsistent Readers displaying distinct gaze patterns associated with comprehension outcomes.

Theoretically, these findings support a cautious interpretation of Schmidt's Noticing Hypothesis and VanPatten's Input Processing framework. The results suggest that visual attention to verb forms is related to more accurate comprehension, but they do not imply that fixation alone equals conscious noticing or acquisition. Rather, eye-tracking data, when interpreted together with comprehension outcomes,

provide evidence of how attention is allocated during form-meaning mapping. The findings also highlight the role of crosslinguistic influence: Mandarin Chinese does not encode tense through verb inflection, and English provides only limited preparation for the richer morphology of Spanish. These prior-language experiences help explain why learners may initially treat Spanish verb forms as secondary or redundant cues.

Pedagogically, the study argues that input should be designed so that learners cannot bypass verb morphology and still arrive at the intended interpretation. Structured input activities, contextualized interpretation tasks, visual enhancement, and reading-strategy support offer practical ways to make verb forms more visible, meaningful, and necessary for comprehension. These recommendations are not intended to replace communicative teaching, but to strengthen it by helping learners process the grammatical information that Spanish encodes through verb morphology.

This study, therefore, illustrates a pathway from eye-tracking evidence to input design. By linking gaze behavior, comprehension outcomes, learner profiles, and instructional strategies, it contributes to research on multilingual processing and offers practical guidance for L3 Spanish teaching. Future research may extend this approach to additional grammatical domains, examine changes in gaze behavior after targeted instruction, and further develop TECS as a resource for tracking how learners' processing strategies evolve over time.

REFERENCES

- Cenoz, Jasone. 2003. The additive effect of bilingualism on third language acquisition: A review. *International Journal of Bilingualism* 7:71-87.
<https://doi.org/10.1177/13670069030070010501>
- Cheng, An Chung. 2002. The effects of processing instruction on the acquisition of *ser* and *estar*. *Hispania* 85:308-323. <https://doi.org/10.2307/4141063>
- De Angelis, Gessica. 2007. *Third or Additional Language Acquisition*. Clevedon: Multilingual Matters.
- Farley, Andrew P. 2005. *Structured Input: Grammar Instruction for the Acquisition-Oriented Classroom*. New York: McGraw-Hill.
- Godfroid, Aline. 2020. *Eye Tracking in Second Language Acquisition and Bilingualism: A Research Synthesis and Methodological Guide*. New York: Routledge.

- Han, ZhaoHong, Eun Sung Park, and Charles Combs. 2008. Textual enhancement of input: Issues and possibilities. *Applied Linguistics* 29:597-618. <https://doi.org/10.1093/applin/amm068>
- Indrarathne, Bimali, and Judit Kormos. 2017. Attentional processing of input in explicit and implicit conditions: An eye-tracking study. *Studies in Second Language Acquisition* 39:401-430. <https://doi.org/10.1017/S027226311600019X>
- Lee, James F., and Bill VanPatten. 2003. *Making Communicative Language Teaching Happen*, 2nd edn. Boston: McGraw-Hill.
- Lee, Sang-Ki, and Hung-Tzu Huang. 2008. Visual input enhancement and grammar learning: A meta-analytic review. *Studies in Second Language Acquisition* 30:307-331.
- Lin, Jo-Wang. 2006. Time in a language without tense: The case of Chinese. *Journal of Semantics* 23:1-53.
- Lu, Hui-Chuan, and An Chung Cheng. 2025. Input processing in the Spanish future tense: Insights from an eye-tracking study. *Foreign Languages and International Studies* 33:39-62. <https://doi.org/10.53106/181147172025060033003>
- Lu, Hui-Chuan, An Chung Cheng, and Joanna Ho. 2025. Verb processing in Taiwanese learners of Spanish: An eye-tracking approach. Paper presented at the 26th International Conference of the Japanese Society for Language Sciences (JSLS). July 12-July 13, 2025, Ehime University, Matsuyama, Japan.
- Lu, Hui-Chuan, An Chung Cheng, and Shu-Ying Hung. 2015. La adquisición del tiempo-aspecto en L3 para los aprendices taiwaneses. *Círculo de Lingüística Aplicada a la Comunicación* 63:200-217. https://doi.org/10.5209/rev_CLAC.2015.v63.50175
- Lu, Hui-Chuan, Li-Chien Kao, Zi-Hao Li, Wen-Hua Lu, and An Chung Cheng. 2024. Corpus of eye movements in L3 Spanish reading: A prediction model. *Asia Pacific Journal of Corpus Research* 5:23-36. <https://doi.org/10.22925/apjcr.2024.5.1.23>
- Rayner, Keith. 2009. The Thirty-Fifth Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology* 62:1457-1506.
- Reichle, Erik D., Keith Rayner, and Alexander Pollatsek. 2003. The E-Z Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences* 26:445-476. <https://doi.org/10.1017/S0140525X03000104>
- Roberts, Leah, and Anna Siyanova-Chanturia. 2013. Using eye-tracking to investigate topics in L2 acquisition and L2 processing. *Studies in Second Language Acquisition* 35:213-235. <https://doi.org/10.1017/S0272263112000861>
- Rothman, Jason. 2011. L3 syntactic transfer selectivity and typological determinacy: The typological primacy model. *Second Language Research* 27:107-127. <https://doi.org/10.1177/0267658310386436>
- Schmidt, Richard. 1990. The role of consciousness in second language learning. *Applied Linguistics* 11:129-158.
- Schmidt, Richard. 2001. Attention. In *Cognition and Second Language Instruction*, ed. Peter Robinson, pp.3-32. Cambridge: Cambridge University Press.

- Sharwood Smith, Michael. 1993. Input enhancement in instructed SLA: Theoretical bases. *Studies in Second Language Acquisition* 15:165-179.
<https://doi.org/10.1017/S0272263100011943>
- Tomlin, Russell S., and Victor Villa. 1994. Attention in cognitive science and second language acquisition. *Studies in Second Language Acquisition* 16:183-203.
- VanPatten, Bill. 1996. *Input Processing and Grammar Instruction: Theory and Research*. Norwood, NJ: Ablex.
- VanPatten, Bill, ed. 2004. *Processing Instruction: Theory, Research, and Commentary*. Mahwah, NJ: Lawrence Erlbaum.
- VanPatten, Bill. 2025. Input processing in adult L2 acquisition. In *Theories in Second Language Acquisition*, 4th edn., eds. Stefanie Wulff, Gregory D. Keating, and Bill VanPatten, pp.64-88. New York: Routledge.
- VanPatten, Bill, and Teresa Cadierno. 1993. Input processing and second language acquisition: A role for instruction. *The Modern Language Journal* 77:45-57.
- Westergaard, Marit, Natalia Mitrofanova, Roksolana Mykhaylyk, and Yulia Rodina. 2017. Crosslinguistic influence in the acquisition of a third language: The Linguistic Proximity Model. *International Journal of Bilingualism* 21:666-682.
<https://doi.org/10.1177/1367006916648859>
- Winke, Paula. 2013. The effectiveness of visual input enhancement on ESL learners' acquisition of grammar. *TESOL Quarterly* 47:355-380.

An Chung Cheng
Department of World Languages and Cultures
University of Toledo
Toledo, Ohio, USA
AnChung.Cheng@utoledo.edu

從眼動追蹤到輸入設計：提升中文學習者西班牙語第三語動詞處理

盧慧娟
國立成功大學

鄭安中
美國托雷多大學

劉子旭
國立成功大學

本研究探討以中文為母語之西班牙語第三語學習者在閱讀時如何處理動詞形態，並說明眼動資料如何應用於輸入設計。研究整合「台灣西語學習者眼動資料庫」(TECS)中三項眼動研究，涵蓋現在／過去時態、未來式及ser/estar + 形容詞結構。結果顯示，學習者常依賴時間副詞、形容詞等詞彙或語境線索，而較少注意承載時態或語意差異的動詞形式；對動詞區域較長的注視通常與較高理解正確率相關。語料層級的預測與分群分析亦顯示不同閱讀型態。本文主張，教學應透過結構化輸入、視覺強化與閱讀策略訓練，使動詞形態更顯著、更具意義，並成為理解不可忽略的線索。

關鍵詞：眼動研究、西班牙語第三語習得、動詞處理、輸入處理、結構化輸入、閱讀型態

STUDENTS' EXPERIENCES WITH GENERATIVE AI IN SOURCE-BASED WRITING: PERCEPTIONS, CONCERNS AND MEDIATIONAL STRATEGIES*

Taimin Tammy Wu
Providence University

Chen-Yu Chester Hsieh[†]
National Tsing Hua University

ABSTRACT

Since the emergence of Generative AI (GenAI), AI-assisted writing has become an increasingly important topic in second language (L2) writing research. However, relatively little is known about how students use GenAI alongside other mediational resources when completing source-based writing (SBW) tasks. Drawing on sociocultural theory and activity theory, this study investigated how 14 Taiwanese L2 university students used GenAI and other mediational resources during English SBW tasks. Students completed three SBW projects supported by ChatGPT, ChatPDF, classroom instruction, and teacher feedback. Data were collected through semi-structured interviews and textual analysis of students' written drafts and revisions. The findings showed that students used GenAI primarily to understand source texts, verify information, organize ideas, and support citation-related decisions. At the same time, students expressed concerns about inaccurate outputs and the reliability of AI-generated content. Beyond GenAI, students relied on multiple mediational resources, including peer discussion, teacher feedback, instructional materials, library databases, and writing center consultation. The findings suggested that AI-assisted SBW is best understood as a mediated activity in which learners coordinate technology, social

* This study was supported by the Ministry of Education, Taiwan. We are grateful to the two anonymous reviewers for their insightful comments. Any remaining errors are our own.

[†] The corresponding author

and institutional resources throughout the writing process.

Keywords: source-based writing, second language writing, Generative AI, sociocultural theory, activity theory, mediational strategies

1. INTRODUCTION

This study investigated the role Generative Artificial Intelligence (GenAI) tools play in second language students' source-based writing (SBW) process, as well as the challenges second language (L2) student writers encounter when using both GenAI and other mediational tools to work on their SBW tasks. In recent years, the rapid development of GenAI has reshaped L2 writing practices. Various GenAI tools offer learners immediate access to linguistic support, idea generation, and feedback, thereby transforming how writing tasks are approached and completed. As these tools become increasingly integrated into educational contexts, understanding their role in shaping writing processes has become an important area of inquiry.

At the same time, SBW, a task that requires students to select creditable sources on an issue, do the readings, synthesize the viewpoints from multiple sources while constructing their own arguments (Tarchi et al. 2025), remains a central yet challenging component of academic literacy development. For L2 writers, this process is particularly demanding, as it involves not only language proficiency but also the ability to manage multiple cognitive and rhetorical demands simultaneously. With the introduction of GenAI tools, the ways in which learners engage with these complex processes are likely to shift in significant but not yet fully understood ways.

Existing research on GenAI in L2 writing has largely focused on its effectiveness in improving writing outcomes, such as grammatical accuracy and overall quality (Link et al. 2024; Kohnke et al. 2025; Shen et al. 2023). While these studies provided valuable insights, they tended to adopt a product-oriented perspective, paying less attention to how learners interact with AI during the writing process. In particular, there is still limited understanding of how GenAI functions as one of multiple mediational resources that learners draw upon in completing complex writing tasks such as SBW.

To address this gap, this study adopted a sociocultural perspective, drawing on sociocultural theory (SCT) and activity theory (AT) to examine how L2 writers engage with GenAI and other mediational resources during source-based writing tasks. Rather than treating GenAI as an isolated tool, this study conceptualizes writing as a mediated activity shaped by the dynamic interaction of tools, social relationships, and institutional structures.

2. LITERATURE REVIEW

2.1 Source-Based Writing in L2 Contexts

Source-based writing (SBW) refers to writing tasks that require learners to select quality readings, read, evaluate, and integrate information from multiple sources to produce a coherent written text (Tarchi et al. 2025). Unlike independent writing, SBW involves complex cognitive and rhetorical processes, including selecting relevant information, organizing ideas, and synthesizing perspectives across texts. These processes are recursive rather than linear, requiring writers to move between reading and writing while constructing meaning. SBW is therefore, recognized as a challenging academic literacy practice for L2 writers. In particular, many L2 writers may have limited prior experience with source attribution and citation practices in academic writing. SBW is also difficult not only because it demands linguistic competence, but because it requires students to coordinate source comprehension, citation conventions, and rhetorical decision-making.

2.2 GenAI in L2 Writing: Affordances and Limitations

Recent developments in AI, particularly GenAI, have significantly influenced L2 writing instruction. AI-powered tools such as ChatGPT, Grammarly, and automated writing evaluation (AWE) systems offer immediate feedback, language support, and personalized learning experiences (Kohnke et al. 2025). These tools can assist learners in

improving grammatical accuracy, expanding vocabulary, and generating ideas, thereby lowering cognitive barriers during the writing process.

Empirical studies have demonstrated that AI-mediated feedback can enhance writing performance, particularly in terms of accuracy and surface-level revisions (Shen et al. 2023). In Siddiqui et al. (2025), AI tools were found to be an effective scaffolding tool enhancing the process-approach in an L2 writing classroom. Students brainstorm, develop argumentation, organize their written content, pay attention to rhetoric, revise and reflect via the help of AI. Students' agency was also high according to the self-reported score. While such findings provided valuable insights into the effectiveness of AI tools, they often overlook how learners engage with these tools during the writing process.

Moreover, existing research tends to treat AI as an isolated instructional tool rather than examining how it interacts with other mediational resources in writing contexts. As a result, relatively little is known about how learners strategically use AI alongside other resources—such as peer collaboration, teacher feedback, and institutional support—when completing complex tasks like SBW. This gap is particularly evident in such contexts, where learners must coordinate multiple sources of information and mediational resources simultaneously.

2.3 A Sociocultural Perspective on L2 Writing: Mediation and Activity Systems

From a sociocultural perspective, writing is not viewed as an isolated cognitive activity, but as a socially mediated process shaped by interactions among learners, tools, and contexts (Lantolf and Thorne 2007). Rather than acting independently, learners rely on cultural artifacts, symbolic tools, and social relationships to regulate their thinking and accomplish writing-related goals. In L2 writing contexts, mediation may occur through various resources, including source texts, teacher guidance, peer interaction, and digital technologies.

Building on this perspective, Lantolf et al. (2020) argued that learning and development occur through different forms of regulation, including object-regulation, other-regulation, and self-regulation. These forms of regulation are particularly relevant in technology-rich writing

environments, where learners must continuously evaluate and coordinate multiple resources while making writing decisions. From this perspective, GenAI should not be viewed simply as a technological aid, but as a mediational artifact that may shape how learners interpret information, construct arguments, and regulate their writing processes.

While sociocultural theory (SCT) emphasizes mediation as the foundation of human learning and development, activity theory (AT) provides a broader framework for understanding how multiple mediational resources interact during goal-directed activities. From an AT perspective, mediation occurs through multiple interconnected elements within an activity system. In writing contexts, learners' actions are shaped not only by tools, but also by rules, communities, and the division of labor that organize participation in writing activities (Engeström 1987, 2001).

Contradictions, among these elements, may emerge when learners encounter tensions between available resources and their writing goals. According to AT, such contradictions are not merely barriers to task completion but potential sources of learning and development, as they may prompt individuals to reorganize their actions and adopt new ways of participating in an activity (Mak and Lee 2014). Rather than being viewed as obstacles, such contradictions may prompt learners to reorganize their actions and negotiate new ways of participating in writing activities. This perspective provides a useful lens for examining how students negotiate the affordances and limitations of GenAI while completing SBW tasks.

2.4 Mediational Strategies in L2 Writing

Research adopting sociocultural perspectives has increasingly examined writing strategies as mediated actions rather than individual cognitive choices (Lei 2008, 2016). From this perspective, L2 writers draw on multiple mediational resources while accomplishing writing tasks, including artifact-mediated strategies such as digital tools and reference materials, community-mediated strategies such as peer collaboration and teacher feedback, and institution-mediated strategies such as writing center consultation and library support (Li 2012; Park and De Costa 2015; Ren and Zhu 2023). These studies showed that writing strategies are

shaped by the affordances and constraints of the surrounding activity system, rather than by writers' individual choice alone.

Importantly, mediational resources rarely function independently. Instead, writers often coordinate multiple forms of support simultaneously while negotiating task demands and writing goals. Lei (2008), for example, argued that writing strategies should be understood as mediated actions embedded within broader sociocultural contexts rather than isolated techniques. Similarly, Ren and Zhu (2023) demonstrated that student writers strategically shifted among different mediational resources during thesis writing, depending on the challenges encountered at different stages of the composing process. Together, these studies suggested that effective writing development depends not on a single resource, but on writers' ability to coordinate multiple mediational resources in response to evolving writing needs.

However, while previous studies have identified various mediational sources that support L2 writing, relatively little attention has been paid to how these resources are coordinated in Gen-AI mediated writing environments. In particular, it remains unclear how GenAI functions as one mediational resource among many others in an L2 writing classroom, and how L2 writers negotiate its use alongside peer interaction, teacher feedback, and institutional support when completing complex SBW tasks.

2.5 Learner Agency in Mediated Writing Contexts

Within sociocultural and activity-theoretical frameworks, agency is not viewed as a fixed personal trait but as a dynamic process that emerges through learners' interactions with mediational resources, opportunities and contextual constraints (Song and Kim 2016; Tigchelaar 2020). Rather than simply following available resources, learners exercise agency by evaluating, adapting, accepting, or rejecting different forms of support while pursuing their writing goals.

Research has shown that L2 writers exercise agency when negotiating feedback, selecting writing strategies, and responding to instructional expectations (Jang 2022). In GenAI-supported writing environments, agency becomes particularly important because learners must make

continuous decisions regarding whether AI-generated suggestions are useful, accurate, or appropriate for their writing purposes.

Recent research further suggested that the educational value of GenAI depends not only on the tool itself, but also on how learners purposefully engage with AI-generated support during the writing process (Siddiqui et al. 2025). However, despite growing interest in GenAI-assisted writing, how learner agency emerges through students' negotiation of AI tools and other mediational resources remain underexplored.

2.6 Gap Statement and Research Questions

While a growing body of research has examined how GenAI tools support L2 writing in terms of vocabulary development, grammatical accuracy, and organizational improvement (Kohnke et al. 2025; Link et al. 2024; Shen et al. 2023), such studies have largely treated GenAI as an isolated instructional tool. From a sociocultural and activity-theoretical perspective, writing is not a tool-driven process, but a mediated activity shaped by the dynamic interaction of multiple elements, including tools, rules, community, and division of labor (Engeström 1987, 2001; Lantolf and Thorne 2007).

Yet, existing research has seldom examined how GenAI operates within this broader mediational system, particularly in relation to other resources such as peer interaction, teacher feedback, and institutional support. To address this gap, the present study investigated how L2 student writers engage with GenAI tools within a broader activity system. Three research questions were addressed:

1. In what ways did the students use GenAI as a mediational tool for source-based writing?
2. What kinds of concerns did the undergraduate participants express when using GenAI tools for this writing task?
3. What other kinds of mediational strategies did the participants adopt to complete their written assignment, after using the GenAI tools?

3. METHOD

3.1 Participants and Context

Participants were 14 sophomore students recruited through criterion sampling (Dörnyei 2007) from the Department of English Language, Literature and Applied Linguistics at a private university in central Taiwan. They were enrolled in Intermediate English Composition I and II during the fall and spring semesters. The participants had intermediate English proficiency (approximately CEFR B1–B2), corresponding roughly to TOEIC (Test of English for International Communication) scores between 550 and 750. Prior to taking these courses, all participants had successfully completed Beginning English Composition I and II during their freshman year. None of the participants had previous experience with SBW before enrolling in Intermediate English Composition I and II, despite having studied English for no less than 12 years.

3.2 Instruments

All participants filled out a background information questionnaire (Appendix A) and participated in semi-structured interviews. The background information questionnaire investigated students' past learning experiences of English writing and SBW. Questions included, for example, "What writing/composition course(s) had you taken before Intermediate English Composition (I) and (II) at university?" and "Had you learned or practiced citation writing before Intermediate English Composition (I) and (II)?"

The semi-structured interviews were designed to explore students' perceptions of GenAI use, their concerns regarding AI-assisted writing, and the mediational resources they drew upon during SBW. The interviews were conducted primarily in Mandarin Chinese, the participants' first language, to encourage more detailed and spontaneous responses. Each interview lasted approximately 20 to 30 minutes. The interview questions (Appendix B) were designed based on the AT framework, centering around artifacts, social interaction (e.g., peers and teachers) and institutional structures. For example, questions such as "How did the guidelines, examples, and resources provided in the class

support your ability to identify the main ideas and supporting details and develop an effective source-based essay?” and “How did discussions with peers or exposure to AI tools shape your understanding of the sources you found and influence the sources you chose to use in your source-based writing?” were designed to elicit participants’ artifact-mediated and peer-mediated strategies.

3.3 Instructional Procedures and Data Collection

During the 2024-2025 academic year, participants did one SBW project in the fall semester and two SBW projects in the spring semester, under the supervision of the same instructor with a doctoral degree in Applied Linguistics and specializing in second language writing. Multi-drafting was applied to the three SBW projects. The first SBW project was a group project practicing a comparison and contrast genre. Participants were asked to write either about whether immigrating to Mars makes life better or staying on Earth or about differences in language use between males and females from a sociolinguistic perspective. The second project and the third SBW projects were individual and group projects, respectively, both requiring students to practice argumentative essays.

For each SBW project, students were provided with grading rubrics, guidelines on selecting credible sources, instructions on using GenAI to identify the main ideas of readings, and access to the Purdue OWL website for learning proper citation formats. The grading rubric consisted of ten criteria, including thesis statement, in-text citations and reference list, organization, logical development, language use, and editing. The rubrics were intended to help students understand the instructor’s expectations and follow the academic writing conventions required for SBW. The instructor incorporated these guidelines into lecture slides. Definitions of credible and non-credible sources were provided, along with citation language, including sentence patterns for in-text citations and guidelines for preparing a reference list based on information from Purdue OWL.

To support reading comprehension, the instructor demonstrated how to write prompts for ChatPDF and guided students in evaluating whether ChatPDF accurately identified the main ideas of the selected sources. During the revision stage, students were asked to indicate sections in

which they had used GenAI for editing or idea development. After submitting the first draft, students participated in a teacher-student conference. A revised draft and a reflection journal were then required one week after the conference. These instructional supports were designed to scaffold students' use of both AI tools and academic conventions throughout the SBW process.

3.4 Data Analysis

Thematic analysis of semi-structured interview data (Seidman 2013) served as the primary analytical approach (Miles et al. 2014; Saldaña 2013). During the first cycle coding stage, concept coding was implemented. According to Miles et al., a concept defines an idea rather than an observable behavior, with the purpose of transferring apparent or tangible objects into a bigger picture, or into “meaning broader than a single item or action” (p. 66). Concept coding was deemed appropriate for the present study because participants often provided lengthy descriptions of their experiences, which could be interpreted and labeled with a conceptual meaning. For example, one participant reported, “The instructor listed all her criteria in the rubric, I can then raise my questions. And the instructor would show us former students’ examples sentence by sentence, so that I know how to write my own draft.” Another participant reported, “The instructor showed us her former students’ writing. I read those writing pieces, and then I borrowed some sentence structures produced by these former students.” These descriptions were coded as “using peer examples.”

Similarly, one of the participants reported, “Compared with the instructor, the instructor in this course interacted with the students more often and cared about how we are doing with our assignment. I found it very helpful.” Another participant noted, “Meeting with the teacher helps me understand the problem and point of my writing better. AI tended to cater to whatever I write. The instructor gave me much more feedback and revision suggestions.” These descriptions were coded as “using teachers’ feedback.”

During the second-cycle coding stage, related codes were grouped into broader categories and themes. For instance, “using peer examples” and

“using teacher feedback” were found to reflect a common pattern of drawing on social support during the writing process. These codes were therefore grouped under the category “community-mediated strategies: distributed support through peers and experts,” which subsequently emerged as one of the major themes in the data. Following this procedure, codes were refined, revised, and organized into themes informed by the AT framework.

Textual analysis of students’ written products was conducted as a supplementary source of evidence to triangulate students’ reported experience and writing practices. The textual analysis focused on students’ use of in-text citations, integration of source materials, and the construction of reference lists. These textual features were examined to better understand how students incorporated external sources into their writing and how they constructed meaning through SBW practices.

4. ROLE AND RULE MEDIATION THROUGH GENAI

Analysis of the interview data revealed that students perceived GenAI as a valuable mediating resource throughout the SBW process. Their reported uses of GenAI can be broadly grouped into two categories: Role mediation and rule mediation. Role mediation refers to instances in which GenAI helped students fulfill the role(s) of a qualified writer in the SBW activity, such as understanding source texts, verifying information, generating ideas, and organizing content.

Rule mediation, in contrast, refers to the ways in which GenAI assisted students in complying with the norms, conventions, and requirements governing academic writing, particularly with regard to citation practices and source use. These findings suggest that students drew on GenAI not only to facilitate the completion of writing tasks but also to navigate the institutional and disciplinary expectations associated with SBW. The following sections illustrate these two forms of mediation through excerpts from the interview data and analyses of participants’ drafts.

4.1 Role Mediation

The interview data suggest that GenAI frequently served a mediating role in helping students accomplish key SBW tasks. Specifically, students reported using GenAI to grasp the main ideas of source texts, check the accuracy of information, and generate and organize content for their essays.

4.1.1 Grasping the gist of readings

“When I was talking about capital punishment issues in Taiwan, I asked ChatGPT to tell me the gist of each source I found. Next, I wrote my own paragraph.” (Interview, Andy)

“I sometimes don’t know how to articulate my understanding of the reading, and I don’t know how to turn it into written sentences. For example, when I was doing the project on ‘different linguistics usage patterns between males and females, I guess I have understood the reading, and I understood the different linguistic patterns used by males and females—from the readings. But I didn’t know how to write that down in my draft. So I asked ChatGPT to help me. ChatGPT would then tell me the content and the gist of that reading...And then, I would know how to describe that different linguistic pattern between males and females in my writing—including the influence these differences have on genders.” (Interview, Tony)

4.1.2 Checking information accuracy

“When we were doing that “Immigrating to Mars” project, we were not sure whether water on Mars is liquid or solid. So we asked ChatGPT and got the answer, that all water on Mars is solid. That helped us decide whether to include a specific piece of information in our draft or not.” (Interview, Andy)

“When doing an argumentative essay, for example, capital punishment issue in Taiwan, I usually read the sources first, and then formed my own thoughts about an issue. And then, I asked ChatGPT if there are any misconceptions or wrong understandings in my thoughts. I took this step

before I went on discussing with my peers because if my thoughts are not logical, I would still get stuck when leading discussions with my peers.” (Interview, Chiang)

According to Andy, Tony and Chiang, ChatGPT served as a mediating artifact at earlier stages in the subjects’ source-based writing process. The subjects, Andy, Tony and Chiang, believed that the role as an author contains multiple responsibilities. The responsibilities included understanding the source information correctly, transferring read information to a written drafts and delivering the source information correctly. GenAI was then used to achieve such purposes, mediating Andy, Chiang and Tony to play the role of an SBW author without errors.

4.1.3 Developing and organizing ideas

“When I was doing my capital punishment project, I asked ChatGPT to offer me different stakeholders’ different views on this issue. There are human rights’ groups, lawyers’ views, those who support capital punishment in Taiwan, and there are editorials from media. I asked ChatGPT to organize those viewpoints for me. This step helped me clarify my own thoughts about capital punishment better.” (Interview, Andy)

“When I needed to put this gender linguistic pattern differences into a magazine article, I also asked ChatGPT to help me find out appropriate headlines...For example, I gave ChatGPT the paragraph I just wrote. And then I had ChatGPT come up with a good headline, or a subtitle.” (Interview, Tony)

“I used ChatPDF at the beginning to grasp the main idea of the reading. Then I found that the main ideas given by ChatPDF and ChatGPT differed. For example, if I wanted to summarize a source or a story, ChatPDF tended to give me a more focused summary, whereas ChatGPT may give me completely different information from the same source” (Interview, Drew).

As students proceeded to the next stage of their SBW, process integrating different perspectives from different sources becomes another

important task for SBW writers. Students define their roles as an author obligated to learn diverse knowledge from multiple sources, and then present the synthesized information to the audience. Giving an appropriate headline, similarly, is a task that should be carried out by an SBW author according to the students' definition. Thus, to mediate the subjects to achieve their SBW object, ChatGPT, serving as the instrument, continued to produce a synthesis of different readings based on students' request. Meanwhile, as authors, students also believed that giving appropriate headlines are author's responsibility, and headlines should convey the most important point in a paragraph. To achieve such an object, the AI tool was used once again to generate suitable headlines. The interaction between the subject and the mediating tools is also encouraged since the subjects are able to share part of their homework responsibilities with the AI tool. In this sense, AI-mediated support was closely aligned with the object of the activity, namely, constructing coherent and well-organized source-based arguments.

Taken together, the aforementioned SBW reading, source selection and drafting process suggested that students writing behaviors can be understood as role-mediated actions, in which the perceived responsibilities of an SBW writer shape and how mediational tools are selected and used. This finding is consistent with Lei (2008), who argued that writing strategies should be understood as mediated actions shaped by writers' goals and available resources.

4.1.4 Evidence of role mediation from draft revisions

To further examine how these role-mediation perceptions were reflected in students' actual writing practices, the first draft and revision were compared. The comparison between the first drafts and revisions of the SBW assignment, focusing on in-text citations, integration of sources, and the list of references, suggests that the use of GenAI helped students improve how they introduced and incorporated sources in such writing tasks, despite some remaining issues. Several patterns can be observed across the texts.

One of the most salient improvements in the revisions is the increase in in-text citations. All of the revised drafts include more citations in the

main text. Another noticeable change in in-text citations is that students make more frequent use of evidential expressions such as *according to*, *the study shows*, and *scientists believe*, as illustrated in (1) to (3), which reflects a wider range of strategies that writers can adopt to engage with source materials.

(1)

a. *Men and women growing up in different cultural backgrounds tend to express emotions differently. For example, in Eastern cultures, emotional expression is often more reserved and restrained compared to Western cultures. People in East culture are less likely to say "I love you" and instead express their emotions through actions, like cooking or showing concern for someone's health. In contrast, people in Western cultures prefer to express their emotions and thoughts more directly.* [first draft by Tony and Drew's group]

b. *According to Khaliq's (2024) study on the use of language by male and female characters in Ibsen's play "A Doll's House", there are differences between males and females in communication priorities. The study shows that males ideologically focus on their status and often engage in conversations that involve lecturing or giving advice.* [revision by Tony and Drew's group]

(2)

a. *The main and most important reason that Mars could be our next home is that it is the most similar planet to Earth in the Solar System. There are several clues and pieces of evidence suggesting that billions of years ago, Mars was even more like Earth, with a denser, warmer atmosphere, rivers, lakes, and possibly oceans.* [first draft by Tim, May, and Rebecca's group]

b. *The most compelling reason Mars could become humanity's next home is its striking similarity to Earth compared to other planets in the Solar System (Borowicz, 2024). Evidence suggests that billions of years ago, Mars had a denser and warmer atmosphere, with rivers, lakes, and possibly oceans, making it even more Earth-like in the past. Studying*

Mars can also deepen our understanding of Earth. Scientists believe that Mars might have undergone a similar evolutionary process to Earth, particularly regarding the history of water, which could provide clues about how life may have once existed there (Neukart,2024).
[revision by Tim, Maggie, and Rebecca's group]

(3)

a. *Another major benefit is the planet would provide tremendous scientific benefits. A permanent presence on the planet would allow for in-depth study of Mars' geology, climate, and atmosphere, offering valuable insights into its history, including whether it ever hosted life. This research could deepen our understanding of planetary science and help us learn more about the potential of life on other planets. In the long term, it could also aid in humanity's broader goals of space exploration.*
[first draft by Sue's group]

b. *Another major benefit is that a permanent presence on Mars could provide tremendous scientific value. Weir (2021) suggests that such an existence will help in-depth study of the geology, climate and atmosphere of Mars, and thus understand its history. According to Weir, it could deepen our understanding of planetary science and potentially contribute to humanity's broader space exploration goals.* [revision by Sue's group]

These changes indicated, more or less, the students' enhanced knowledge of, and ability to adopt, English academic writing conventions, where claims are expected to be explicitly attributed to sources rather than presented as general statements.

4.2 Rule Mediation

According to the interview data, students frequently used GenAI to navigate the conventions and requirements of SBW. They reported consulting GenAI to check citation formats and source attribution practices. These reported uses were further corroborated by revisions identified in their drafts.

4.2.1 Checking citation formats

“When I was doing my capital punishment project, I needed to read some court judgement documents, and I referred to Blue Book version 21st because I needed to cite legal documents. I first pasted that source website in ChatGPT and asked it how I should cite this document. Unfortunately, ChatGPT cited the author’s name initial wrong. The rest parts seemed to be ok...And...I found out AI’s errors by checking with my instructor in class.” (Interview, Tony)

In this quotation, Tony described several steps where he used AI to mediate his SBW tasks. First, being aware of the role an author should play, Tony checked Purdue OWL and learned that he should use legal documents from court websites to support his argumentation on capital punishment.

When encountering incorrect citation output generated by ChatGPT, Tony did not passively accept the AI-generated response. Instead, he cross-checked the citation format with his instructor, demonstrating his orientation toward academic writing conventions as rule-mediated resources. This process also reveals a moment of contradiction within the activity system, where the mediating tool, AI, failed to align with the established rule of academic writing. Such contradiction did not hinder Tony’s writing process. Rather, it prompted further mediation, as Tony actively sought additional human support to resolve the inconsistency. Tony’s quotation suggested that rule-mediation in AI-supported writing is not a linear process, but a dynamic negotiation between tools, academic writing conventions, and available social resources in an EFL classroom.

4.2.2 Evidence of rule mediation from draft revisions

Additional evidence for rule mediation can be found in the textual analysis of students’ revisions. The revised texts also demonstrated improvements in both the number and formatting of references. Several students added more sources and revised their reference entries to better align with APA style by including more complete information about

authors and publications. Some students also incorporated a greater number of higher-quality sources, such as journal articles and more recent publications. Figures 1 and 2 illustrate these improvements.

References:

n-situ resources for infrastructure construction on Mars by J Liu

<https://www.sciencedirect.com/science/article/pii/S204604302100006X>

Why go to Mars?

https://www.esa.int/Science_Exploration/Human_and_Robotic_Exploration/Exploration/Why_go_to_Mars

What do Mars and Earth have in common?

<https://www.britannica.com/question/What-do-Mars-and-Earth-have-in-common>

Florian Neukart. (29 February 2024) Towards sustainable horizons: A comprehensive blueprint for Mars colonization

<https://www.sciencedirect.com/science/article/pii/S2405844024022114>

The Habitability of Mars

<https://www.marssociety.ca/the-habitability-of-mars/>

Figure 1. Reference List in the First Draft by Tim's Group

References

- Brodowicz, M. (2024). Advantages And Disadvantages Of Moving To Mars. *Aithor*.
<https://doi.org/https://aithor.com/essay-examples/advantages-and-disadvantages-of-moving-to-mars/>
- Liu, J., Li, H., Sun, L., Guo, Z., Harvey, J., Tang, Q., Lu, H., & Jia, M. (2022). In-situ resources for infrastructure construction on Mars: A review. *International Journal of Transportation Science and Technology*, 11(1), 1-16.
<https://doi.org/https://doi.org/10.1016/j.ijtst.2021.02.001>
- Neukart, F. (2024). Towards sustainable horizons: A comprehensive blueprint for Mars colonization. *Heliyon*, 10(4).
<https://doi.org/https://doi.org/10.1016/j.heliyon.2024.e26180>
- Plant-Weir, E. (2021). The Habitability of Mars. *The Mars Society of Canada*.
<https://doi.org/https://www.marssociety.ca/the-habitability-of-mars/>
- West, D. M. (2020). Five reasons to explore Mars. *Brookings*.
<https://doi.org/https://www.brookings.edu/articles/five-reasons-to-explore-mars/>

Figure 2. Reference List in the Revised Draft by Tim's Group

5. CONTRADICTIONS AND AGENCY IN AI-MEDIATED WRITING

While AI tools mediated L2 student writers' SBW process, students also have concerns when AI did not seem to meet the goals students originally expected to achieve for their assignment. For example, Rebecca talked about the challenges she encountered when using AI to mediate her revision process. Sue found that the readings generated by AI did not match her drafting directions. Drew, who did a lot of information-check between sources generated by GenAI and sources from other scholarly search engines, found inconsistent information between these two sets of platforms.

5.1 Agency Mediation

“For project#2, the eco-tourism assignment, I did it mostly on my own. And for project#3 capital punishment, I had AI help me with only the last paragraph. I did four rounds of revisions on my project #2. If AI could have revised my paragraphs better, I would not have had to do the revision for four times. I wasn’t sure whether the revision generated by AI meet the instructor’s expectations. So I feel like sometimes AI doesn’t really help me too much. Maybe I should give AI clearer prompt, so that AI would give me the real answer I want. Otherwise, I got rubbish answers” (Interview, Rebecca)

“When I was doing my overtourism project, I wanted to explain how a large number of tourists’ visits in Venice lowered the level of the sea water of the city. I did a lot of readings, but I found many readings not satisfying. The content did not match my purpose. For example, I wanted to focus on how human errors and human damages had affected Venice. But the reading I found from Google talked about pollution in Venice. This is not what I wanted. To solve this problem, I told ChatGPT, ‘this is what I am currently writing, and I want to focus on human impact on Venice. Please recommend me more relevant readings that focus on human errors and damages in Venice.’ (Interview, Sue)

5.2 Contradictions between AI and Other Mediating Tools

“The extent to which I decide to use answers from ChatGPT depended on how difficult the words it generated for me. If I found the words too difficult, too deep or has multiple meanings, I would change the hard words into an easier one.” (Interview, Drew)

“When I was writing my project#3 on whether DNA affects the way males and females communicate, I found a particular DNA term generated by ChatGPT differed from the one I found in another reading. I asked AI to explain, whether human being’s DNA influence communication between different genders? And then...AI seemed to generate different answers and different terms. For example, AI generated terms like ‘basal ganglia’ and ‘corpus striatum,’ but I wasn’t sure if ‘basal ganglia’ and ‘corpus striatum’ actually referred to, or equal to the concept of “DNA”

mentioned in the readings I encountered in the literature. Since my readings from the literature did not include these two terms, I ended up didn't check whether 'basal ganglia' and 'corpus striatum' are accurate or not." (Interview, Drew)

The findings of research question 2 further revealed that students' concerns regarding AI use can be understood through the lens of contradiction within the activity system. According to AT, contradictions, manifested as tensions or inconsistencies among mediational elements, serve as a driving force for development rather than merely obstacles (Engeström 1987; Ren and Zhu 2023).

In the present study, students' experiences with AI frequently involved such contradictions. For instance, Rebecca expressed dissatisfaction with AI-mediated revisions, pointing out that AI did not do satisfactory revisions, so she eventually revised the draft on her own. Rebecca questioned whether AI-generated outputs aligned with the instructor's expectations. Rather than fully relying on AI, she began to reconsider how to formulate more precise prompts, indicating an emerging awareness of how to regulate her own use of AI as a mediational tool.

A similar form of contradictions can be observed in Sue's experience. While searching for relevant sources for her overtourism project, Sue found that the materials she initially located did not align with her intended writing focus. The discrepancy between her writing goals, or her object, and the available resources created the tension within the activity system. To resolve this contradiction, Sue actively reformulated her query to ChatGPT by specifying her writing purpose and desired focus. This shift demonstrated that Sue did not passively consume information provided by external sources. Instead, she strategically reoriented the mediational tool to better serve her writing objectives (Mak and Lee 2014).

Drew's case further illustrated how contradictions may arise between different mediational resources (Mak and Lee 2014). When encountering inconsistent terminology between AI-generated responses and academic reading, Drew hesitated to adopt unfamiliar terms and instead simplified or avoided them. Although this decision reflects uncertainty, it also indicated an evaluative stance toward competing sources of information.

Drew's decision also suggested that students are not merely dependent on AI outputs, but are engaged in ongoing decision-making processes.

Aforementioned moments of tension did not lead to disengagement. Instead, they prompted students to actively adjust their strategies, including refining prompts, re-evaluating sources, or selectively adopting AI-generated content. This process can be conceptualized as agency mediation, in which learners actively negotiated and orchestrate available mediational resources to achieve their writing goals. From a sociocultural perspective, agency is not an inherent individual trait but emerges through interactions with tools, rules, and social contexts (Song and Kim 2016). In this sense, the contradictions encountered by students in AI-mediated writing environments may serve as critical sites for the development of agency, as learners are compelled to make informed decisions about how to use, adapt, or resist technological assistance. This aligns with recent work, suggesting that learners' agency in AI-supported writing is shaped by how they purposefully engage with AI as one of multiple mediational resources (Siddiqui et al. 2025).

6. MULTIPLE MEDIATIONAL RESOURCES BEYOND GENAI

In addition to AI tools, students also reported drawing on a range of other mediational sources to support their composing processes. These resources can be organized into three main categories: Community-mediated strategies, artifact-mediated strategies, and institution-mediated strategies.

6.1 Community-Mediated Strategies: Distributed Support through Collaboration and Expert Feedback

As illustrated in the following interview excerpts, students frequently engaged in community-mediated strategies, in which writing processes were supported through interactions with peers and instructors.

6.1.1 Collaborating with peers

“For our collaborative project, we usually got together and discussed the organization of our paragraphs. We also went over the sources everyone found, and decided which would be the main reading(s) we wanted to cite.” (Interview, Rebecca)

“I usually had Andy take a look at the paragraphs I drafted. He sometimes commented on my choice of words, and I would follow his suggestions. For example, one time Andy fixed my sentences. He changed my original sentences into ‘instead of rushing to abolish [capital punishment]... I couldn’t have come up with a phrase like this if I did the paragraph on my own.’” (Interview, Rebecca)

“We did talk about our citations. If I comment on a citation on my own, I am not sure if I understand the reading correctly. I can’t guarantee that my peers would agree with me. So what our group did, was that we got together and discussed those readings.” (Interview, Tim)

6.1.2 Using teacher’s feedback from student-teacher conference

“When doing our collaborative project on capital punishment in Taiwan, we talked about our draft with the instructor at the student-teacher conference. The instructor explained paragraph by paragraph... Sometimes I got stuck at a certain point in my argument, and didn’t know how to move on. But with the instructor’s feedback, we were better at knowing how to argue and where the key point lies in. Comparatively, AI tools tended to support our argument rather than giving us constructive feedback.” (Interview, Chiang)

“For our collaborative project on capital punishment in Taiwan, we didn’t know how to structure the essay at the very beginning. We seemed to repeat the same information in paragraph one, two and three, and we didn’t know how to do transition. The instructor gave us a lot of suggestions, on how to organize.” (Interview, Andy)

“When I was doing my summary writing, I didn’t have much experience writing this kind of genre before. I didn’t know how to begin.

At the student-teacher conference, the instructor commented on my first draft and told me where to revise, and how to make my point clear for the audience. And the, I learned how to start my first paragraph and how to give a conclusion.” (Interview, Drew)

The interactions mentioned in these excerpts illustrated how writing responsibilities were distributed across members of the learning community, reflecting division of labor within the activity system.

6.2 Artifact-Mediated Strategies: Instructional Tools as Scaffolding for SBW Development

In addition to social support, students relied heavily on artifact-mediated strategies, particularly instructional materials provided in class.

6.2.1 Using in-class materials

“The instructor’s rubric helped. For example, there were different items in the rubric, such as organization, grammar, editing and so on. By looking at the break-down scores in the rubric, I know which aspect of my writing needs improvement. There were also good examples from former students. The instructor presented those examples in class, and went sentence by sentence. She showed us how those examples had done a good job on transitions, logic and content. Then I had a better idea of how to write this project.” (Interview, Melissa)

“I usually started my writing process by looking at the good examples our instructor provided us. Of course I wouldn’t copy them, but I would know how other students talk about this topic. And the, I read the instructor’s rubric, and then I began to write.” (Interview, Tim)

“When doing that capital punishment project, I watched that YouTube video uploaded by our instructor. That video debated the issue of whether CP should be abolished in Taiwan. The video presented different views of this issue, and there were huge differences among different stakeholders.” (Interview, Andy)

“I used the instructor’s PowerPoint slides to write my project#2 and my reflection journals. The ‘they say, I say’ formula listed in the slides were very helpful when I needed to cite. I followed this formula, used APA style and then I used AI tools to check my grammar. Also the “move1-move2-move 3” formula guided me to reflect my composing process.” (Interview, Sue)

“I read the good examples done by the instructor’s former students, finding that some good examples did include a lot of references. To me, many references in an example means that the author worked very hard. That encouraged me and I thought I should also work hard by finding as many references as possible. I also looked at how those former students organize their essay, and tried to mimic their structures when doing my own argumentation. This step helped me express my main ideas more clearly.” (Interview, Tony)

“I honestly didn’t know how to begin my argumentative essay when doing project#2 “ecotourism, until I read the six good examples uploaded by the instructor. And then I learned that ok, I need to analyze an issue for this project. I realized that an author needs to first read a lot so that he/she knows how to approach an issue, and then how to write.” (Interview, Rebecca)

6.2.2 Utilizing in-class activities

“I remember that we had an in-class activity, in which we were required to read three English reading on tourism. And then we needed to synthesize the information from the three readings and write a paragraph. The activity pushed me to read a lot, and re-describe the information in my own words. I needed to re-read them back and forth for many times, and found useful points.” (Interview, Melissa)

“When we were doing activity 1 and 2 in class, the instructor provided us three readings. Those readings had clear themes...When I was doing project #2 on my own, I felt like my draft content was bad. I didn’t know

how to synthesis information from different readings either. Then it occurred to me that I have done this read-to-write activity in class. I eventually picked up one reading provided by the instructor from activity 1. Later, I knew how to write my own draft.” (Interview, Edison)

These artifacts functioned as mediational tools that guided students’ understanding of genre conventions and supported the development of their source-based writing.

6.3 Institution-Mediated Strategies: Extending Mediation beyond the Classroom

Beyond classroom-based mediation, students also drew on institutional resources to support their writing processes. Library workshops and writing center consultation were the two institution-mediated strategies students relied on when carrying out their SBW task.

6.3.1 Library sources

“When working on our first SBW collaborative project, I used that ‘jumper’ database on our university library website. I downloaded an article talking about gender roles. I read the article, and if I found that the article content matched the goal of our draft, I would cite it.” (Interview, Tony)

“The first time I was trying to do SBW, I found citation difficulties. And then our instructor told us the library offers EndNote workshop.” (Interview, Tim)

“When I was working on my project, I was talking about how Gaomei Wetlands might be negatively affected by overtourism. There was very limited useful information on the Internet, so I consulted EndNote. And then I found that EndNote quickly identified the correct sources I wanted—with accurate citation format and very scholarly reading content. The source has clear author’s name, and there was a huge amount of it.” (Interview, Tim)

6.3.2 Writing center consultation

“When I was doing my summary writing assignment, the instructor recommended us to visit the writing center, where another professor would help me. I brought my draft and the professor commented on my writing, giving me suggestions. I found this very effective because the writing center professor did not give me very difficult suggestions. He was able to, and willing to follow *my* English writing proficiency level, and offer me constructive and appropriate feedback. I found his suggestions very helpful for my revision.” (Interview, Tim)

These multiple mediational resources were also reflected in students’ revised drafts. Compared with the first drafts, the revised texts contained more detailed information from source materials and demonstrated a wider range of source integration strategies, as illustrated in Examples (4) to (6). As shown in Example (4), the revised texts incorporated additional source materials and more detailed supporting information, resulting in longer and more elaborated discussion.

(4)

- a. *Conversation is something we do every day in our daily life. It can convey messages and bring each other closer. However, the content of conversations between men and women also differs. According to the study by Bischooping (1993), research shows that women talk to women about relationships, emotions, and family. Most of the conversations between men are related to money, entertainment, or business. Women usually share personal experiences in conversations, expecting to receive empathy and support, emphasizing caring and emotional communication. And focus on maintaining relationships. In contrast, men’s dialogues tend to focus on objective facts, reasoning, and conclusions, showing strong practicality and logic. These ways of communicating highlight the different expectations of each gender and influence their interactions in relationships and the workplace. [first draft by Tony and Drew’s group]*

- b. *In addition to the ideological focus on gendered dialogue and the communication goals between men and women, conversation styles for males and females may also differ. According to Brodowicz (2024), female chat content often focuses on building and maintaining relationships, including sharing personal experiences, emotions, and details of daily life. This communication style emphasizes emotional connection and mutual support to foster social bonds. Additionally, women tend to use more nonverbal cues, such as facial expressions and body language, to convey meaning in conversations. In contrast, male chat content typically centers on sharing information, solving problems, or expressing opinions, with a focus on topics like work, technology, sports, or competitive interests. They rely more on verbal communication, using direct language to express their thoughts and feelings. Overall, males and females have different communication patterns, which are influenced by many factors including socialization, cultural norms, and personal personality traits. Research shows that females tend to use more collaborative language, seek other's recognition and support, and focus on building rapport in conversations. They are more likely to build relationships through conversation, discussing emotions and personal experiences. Males often use more direct and assertive language, prioritize problem-solving, and emphasize status and dominance in interactions. They often view conversations as a tool to negotiate power and focus on accomplishments or facts.* [revision by Tony and Drew's group]

While the content of the original sources was primarily presented through summary or paraphrasing, some of the revisions demonstrated a wider range of strategies, such as the use of verbatim quotations exemplified in (5) and (6).

- (5) *According to Lieberman (2000), encoding is defined as "the process of translating a mental state into an externally visible signal like a facial expression" (p. 111). Decoding, on the other hand, "involves drawing inferences with little effort or attention about the internal mental states and dispositions of other individuals on the basis of subtle sequences of nonverbal cues" (p. 111).* [revision by Tony and Drew's group]

- (6) *Moreover, the extreme cold further complicates life on Mars. "It's not reasonable because it's so cold," Chris (2022) explains.* [revision by Sue's group]

The findings of RQ3 further extended the analysis by illustrating how students' writing processes were mediated through multiple interconnected resources, including community-, artifact-, and institution-mediated strategies. First, community-mediated strategies were evident in students' interactions with peers and instructors. As reflected in the data, peer discussion provided a space for students to exchange ideas, clarify arguments, and negotiate meaning during the composing process. In addition, instructor feedback played a crucial role in guiding students' revisions, particularly in helping them refine their arguments and align their writing with task expectations. These forms of social mediation supported students in making more informed decisions throughout their writing process.

Second, artifact-mediated strategies were prominently observed in students' use of digital tools, including AI tools and online resources. Students reported using Gen AI tools to generate ideas, summarize readings, and revise language forms, while search engines and online materials helped them locate relevant information. These artifacts functioned as cognitive supports that facilitated various stages of writing, particularly in drafting and revising. However, students also noted limitations of these tools, such as inaccurate citations or overly general feedback, which required them to critically evaluate and selectively adopt the suggestions provided.

Furthermore, institution-mediated strategies played a distinct role in supporting students' SBW development. Unlike general online tools, institutional resources such as library databases, EndNote workshops, and writing center consultations provided access to more authoritative and academically appropriate support. For example, library databases enabled students to locate more credible and discipline-relevant sources, while EndNote assisted them in managing citations more systematically.

In addition, writing center consultations offered personalized and level-appropriate feedback that was responsive to students' specific

writing needs. Compared with GenAI-generated feedback, which students sometimes perceived as vague or unreliable, writing center support was viewed as more targeted and pedagogically meaningful. These findings suggested that institutional resources not only supplemented classroom instruction, but also expanded the mediational space of writing by connecting students to broader academic support systems.

When considered as a whole, these findings suggested that GenAI-mediated SBW is best understood as a dynamic activity system in which multiple forms of mediation interact. This interpretation aligns with Lei (2008) and Ren and Zhu (2023), who conceptualized writing strategies as mediated actions embedded in holistic activity systems. It also resonates with Mak and Lee (2014), Lantolf and Thorne (2007), and Lantolf et al. (2020), who emphasized the interconnected roles of feedback, pedagogy, and learning in writing contexts, as well as with sociocultural perspectives that view development as emerging through learners' interactions with tools, social relationships, and institutional support.

These findings also support Beach et al.'s (2016) argument, that writing development should be understood through students' participation in multiple social practices situated within specific instructional and institutional contexts. In the present study, students did not rely on a single tool, but coordinated GenAI tools, peer interaction, teacher feedback, and institutional resources throughout the SBW process. In this sense, learner agency in AI-supported writing emerges not from the individual or the tool alone, but from learners' ongoing negotiation of multiple mediational resources within specific sociocultural contexts.

7. PEDAGOGICAL IMPLICATIONS

The findings of this study suggested that GenAI should not be treated as an autonomous writing assistant, but as a mediational resource that requires explicit instructional guidance. As demonstrated in RQ1 and RQ2, students used GenAI to understand source texts, organize ideas, and support citation-related decisions. However, they also encountered limitations, including inaccurate information, inappropriate suggestions, and outputs that did not align with

their writing goals. These findings indicated that instructors should not only introduce GenAI tools into writing classrooms, but also help students develop critical AI literacy, including how to formulate effective prompts, evaluate AI-generated content, and make informed decisions about when and how GenAI support should be adopted.

The findings also highlighted the importance of designing writing instruction as a multi-layered mediational environment. In addition to GenAI, students relied on peer discussion, teacher feedback, instructional materials, library databases, and writing center consultations throughout the SBW process. These resources complemented one another and supported students at different stages of writing. Rather than viewing Gen AI as a replacement for existing instructional support, writing instructors may consider integrating GenAI into a broader network of mediational resources that encourages students to coordinate technological, social, and institutional forms of support while completing complex source-based writing tasks.

Overall, the findings suggested that effective Gen AI-supported writing instruction depends not only on access to technological tools, but also on the availability of diverse mediational resources that support students' development as independent writers.

8. CONCLUSION

This study examined how L2 student writers used GenAI and other mediational resources while completing source-based writing tasks. The findings suggest that GenAI functioned as one mediational resource among many others, supporting students' reading, drafting, and citation-related decisions. At the same time, students' writing processes were shaped by their ongoing interactions with peers, instructors, instructional materials, and institutional resources. From a sociocultural perspective, AI-assisted source-based writing is best understood as a mediated activity situated within a broader activity system rather than as the use of a single technological tool. As the study was conducted in a single EFL writing course, future research may examine similar processes across different instructional contexts.

REFERENCES

- Beach, Richard, George Newell, and Jennifer VanDerHeide. 2016. A sociocultural perspective on writing development: Toward an agenda for classroom research on students' uses of social practices. In *Handbook of Writing Research*, eds Charles A. MacArthur, Steve Graham, and Jill Fitzgerald, pp. 88–101. New York: Guilford Press.
- Dörnyei, Zoltán. 2007. *Research Methods in Applied Linguistics*. Oxford: Oxford University Press.
- Engeström, Yrjö. 1987. *Learning by Expanding: An Activity-Theoretical Approach to Developmental Research*. Helsinki: Orienta-Konsultit.
- Engeström, Yrjö. 2001. Expansive learning at work: Toward an activity theoretical reconceptualization. *Journal of Education and Work* 14(1):133–156.
- Jang, Jinsil. 2022. An exploratory study on learner agency and second language writing practices of Korean high school students. *Asian-Pacific Journal of Second and Foreign Language Education* 7:31. <https://doi.org/10.1186/s40862-022-00158-1>
- Kohnke, Lucas, Di Zou, and Fan Su. 2025. Exploring the potential of GenAI for personalised English teaching: Learners' experiences and perceptions. *Computers and Education: Artificial Intelligence*, 8:100371. <https://doi.org/10.1016/j.caeai.2025.100371>
- Lantolf, James P., Matthew E. Poehner, and Steven L. Thorne. 2020. Sociocultural theory and L2 development. In *Theories in Second Language Acquisition: An Introduction*, eds Bill VanPatten, Gregory Keating, and Stefanie Wulff, pp. 223–247. London: Routledge.
- Lantolf, James P., and Steven L. Thorne. 2007. Sociocultural theory and second language acquisition. In *Theories in Second Language Acquisition: An Introduction*, eds Bill VanPatten and Jessica Williams, pp. 201–224. Mahwah: Lawrence Erlbaum Associates.
- Lei, Xiao. 2008. Exploring a sociocultural approach to writing strategy research: Mediated actions in writing activities. *Journal of Second Language Writing* 17(4):217–236. <https://doi.org/10.1016/j.jslw.2008.04.001>
- Lei, Xiao. 2016. Understanding writing strategy use from a sociocultural perspective: The case of skilled and less skilled writers. *System* 60:105–116. <https://doi.org/10.1016/j.system.2016.06.006>
- Li, Mimi. 2012. Mediated processes in writing for publication: Perspectives of Chinese science postdoctoral researchers in America. *The Asian ESP Journal* 8(1):117–140.
- Link, Stephanie, Robert Redmon, Yaser Shamsi, and Martin Hagan. 2024. Generating genre-based automatic feedback on English for research publication purposes. *CALICO Journal* 41(1):1–28. <https://doi.org/10.1558/cj.26273>
- Mak, Pauline, and Icy Lee. 2014. Implementing assessment for learning in L2 writing: An activity theory perspective. *System* 47:73–87. <https://doi.org/10.1016/j.system.2014.09.018>

- Miles, Matthew B., A. Michael Huberman, and Johnny Saldaña. 2014. *Qualitative Data Analysis: A Methods Sourcebook*. Thousand Oaks: SAGE.
- Park, Ji-Hyun, and Peter De Costa. 2015. Reframing graduate student writing strategies from an Activity Theory perspective. *Language and Sociocultural Theory* 2(1):25–50. <https://doi.org/10.1558/1st.v2i1.24977>
- Ren, Beibei, and Wei Zhu. 2023. A Chinese EFL student's strategies in graduation thesis writing: An activity theory perspective. *Journal of English for Academic Purposes* 61:101202. <https://doi.org/10.1016/j.jeap.2022.101202>
- Saldaña, Johnny. 2013. *The Coding Manual for Qualitative Researchers*. Thousand Oaks: SAGE.
- Seidman, Irving. 2013. *Interviewing as Qualitative Research: A Guide for Researchers in Education and the Social Sciences*. New York: Teachers College Press.
- Shen, Chen, Penghai Shi, Jirong Guo, Suyun Xu, and Jiwei Tian. 2023. From process to product: Writing engagement and performance of EFL learners under computer-generated feedback instruction. *Frontiers in Psychology* 14:1258286. <https://doi.org/10.3389/fpsyg.2023.1258286>
- Siddiqui, Momin N., Vryan Feliciano, Roy Pea, and Hari Subramonyam. 2025. AI in the writing process: How purposeful AI support fosters student writing. arXiv. <https://doi.org/10.48550/arXiv.2506.20595>
- Song, Bongsun, and Tae-Young Kim. 2016. Teacher (de)motivation from an activity theory perspective: Cases of two experienced EFL teachers in South Korea. *System* 57:134–145. <https://doi.org/10.1016/j.system.2016.02.006>
- Tarchi, Christian, Alessandra Zappoli, Lidia Casado Ledesma, and Eva Wennås Brante. 2025. The use of ChatGPT in source-based writing tasks. *International Journal of Artificial Intelligence in Education* 35:858-878. <https://doi.org/10.1007/s40593-024-00413-1>
- Tigchelaar, Magda. 2020. 'I do the peer review by myself': A Chinese international student's approach to process writing. *Writing and Pedagogy* 12(2-3):396-422. <https://doi.org/10.1558/wap.20355>

Chen-Yu Chester Hsieh
 Department of Foreign Languages and Literature
 National Tsing Hua University
 Hsinchu, Taiwan 300044, ROC
 chesterhugues@gmail.com

APPENDIX A

Background information questionnaire: AI and learning of writing questionnaire

Please answer the following questions to the best of your knowledge.
Check in the blanks when necessary.

1. Your name: _____
2. Year of birth: _____
3. Gender: _____ Male _____ Female
4. Major: _____
5. Native language: _____
6. Second language: _____
7. How many years have you been learning English? _____
8. How many years have you been learning English writing? _____
9. Have you lived in an English-speaking country?
_____ Yes (Go to question 9a and 9b) _____ No (Go to question 10)
- 9a: Which country did you live in? _____
- 9b: How long did you live in that country? _____ months
_____ years
10. What writing/composition course(s) have you taken before Intermediate English Composition (I) and (II) in university/college?
Please list all the course titles. _____
- 11a Have you learned/practiced citation writing before Intermediate English Composition (I) and (II)?
_____ No, this is my first time practicing citation writing.
_____ Yes (Go to question 11b)
- 11b. When did you learn about/ practice citation writing in your learning journey? (Check multiple items if applicable)
_____ Chinese class(es) in elementary school
_____ English class(es) in elementary school
_____ Chinese class(es) during junior high or senior high school
_____ English class(es) during junior high or senior high school
_____ English composition course(s) in university. Please write the course title: _____

____ Other courses in university: Please write the course title

12. Before taking Intermediate Composition (I) and (II), have you used any AI tools (for example, ChatGPT, ChatPDF, Notebook LM or anything else) to help with your own writing homework?

____ No, this is my first time using AI tool(s) in my English composition courses.

____ Yes (Please list the name of that AI tool): _____

APPENDIX B

Interview questions for source-based writing strategy use

1. Can you share any instances where working with your classmates or receiving feedback from your teacher helped you effectively summarize the sources you found?
2. How did the guidelines, examples, and resources provided in the class support your ability to find the main ideas and supporting details and develop an effective source-based writing/argumentation?
3. Were there any specific activities or tasks that you found helpful in understanding and evaluating the reading/sources for your source-based writing? Can you describe how they contributed to your learning?
4. How did the social and cultural context (micro and macro), such as discussions with peers or exposure to AI tools, shape your understanding of the sources you found and influence the sources you chose to form in your summary writing/argumentation?
5. Can you share any challenges you faced when understanding different readings from various sources before doing your summary writing/argumentation? How did you address these challenges?
6. What are the techniques and strategies taught in the class that you still remember? Did the writing techniques and strategies taught in the class help you effectively comprehend your reading and synthesis useful points to produce a summary writing/argumentation? Can you provide an example?

7. How did the class activities, such as group discussions or collaborative projects, contribute to your ability to critically engage with source material and construct a comprehensive source-based writing in your argumentation ?
8. What types of resources or tools, such as scholarly articles, books, online databases or AI tools, did you find most useful in decoding information and supporting your claims in your source-based writing argumentation?
9. Looking back on your experience in the source-based writing projects/argumentation, what do you think helped you the most in accurately presenting the author's viewpoints through the use of ChatPDF and ChatGPT?
10. How do you think the activities and interactions in the class influenced your development of source-based writing/argumentation skills overall?

生成式人工智慧輔助來源整合寫作：學生知覺、疑慮與中介策略

吳黛敏
靜宜大學

謝承諭
國立清華大學

儘管生成式人工智慧輔助來源整合寫作在二語寫作教學上日益受到重視，但相關文獻仍相對有限，本研究基於社會文化理論與活動理論，探討 14 位臺灣英語學習者結合生成式人工智慧及其他中介資源完成來源整合寫作任務的經驗。根據半結構式訪談及學生作業之文本分析，研究發現，學生主要利用生成式人工智慧協助理解來源內容、查核資訊、組織觀點及處理文獻引用問題，但亦對其生成內容的正確性與可信度抱持疑慮。此外，學生亦廣泛運用教師回饋、同儕討論、課堂教材、圖書館資料庫及寫作中心諮詢等中介資源。結果顯示，人工智慧輔助的來源整合寫作應視為一種多重中介活動，學生需協調科技、社群與校內支持資源，以完成寫作任務。

關鍵字：來源整合寫作、第二語言寫作、生成式人工智慧、社會文化理論、活動理論、中介策略

GUIDED USE OF GEN-AI FOR RESEARCH ARTICLE READING: A CASE STUDY OF EFL STUDENTS IN A TAIWANESE MA-TEFL PROGRAM*

Chia-Hui Chiu
Tunghai University

ABSTRACT

Little research has examined instructional approaches for helping EFL students comprehend research articles. To address this gap, this case study investigated the effects of combining Move Analysis instruction and generative artificial intelligence (Gen-AI) prompting instruction on the academic reading comprehension of Taiwanese EFL students enrolled in an MA-level course. Twelve students received instruction and then read two research articles in a counterbalanced crossover design, alternating between ChatGPT (dialogic) and NotebookLM (document-grounded) across tasks. Their reading comprehension, perceived usefulness of the instruction and AI tools, and interaction behaviors were subsequently examined. Results indicated that NotebookLM yielded higher reading comprehension scores, whereas ChatGPT was perceived as slightly more useful; however, neither difference was statistically significant. The Move Analysis instruction was rated as helpful. Participants primarily used ChatGPT to seek clarification and ask questions, whereas NotebookLM was used mainly through its built-in features such as summaries and podcasts. The findings suggest that NotebookLM may offer more structured support for comprehension and synthesis, whereas ChatGPT may facilitate clarification and on-demand understanding.

Keywords: Move Analysis, Gen-AI, research article reading comprehension

* The author would like to thank all students in this study for their participation and engagement.

1. INTRODUCTION

The ability to read research articles in English is essential for academic success in graduate education. However, such texts are often perceived as challenging, particularly by English as a Foreign Language (EFL) readers, due to their dense terminology, complex theoretical constructs, extended argumentation, and intricate discourse structures. These features can impose a substantial cognitive load, thereby hindering comprehension. These challenges are especially relevant in the Taiwanese MA-TEFL context, where students are frequently required to read English-medium research articles for coursework, seminar discussions, and thesis preparation. Yet many enter graduate programs with limited experience reading empirical research articles and may struggle to meet the academic demands. Thus, exploring pedagogical approaches that can facilitate research article comprehension among Taiwanese EFL students may provide valuable insights for higher education and academic literacy development.

From the perspectives of Cognitive Load Theory (CLT; Sweller 2011) and Vygotsky's (1978) concept of the Zone of Proximal Development (ZPD), Gen-AI may support EFL readers by reducing unnecessary cognitive demands through features such as information simplification and immediate explanations while simultaneously providing scaffolded assistance for tasks beyond learners' current independent abilities. These affordances suggest that Gen-AI has the potential to make reading research articles more manageable and interactive. Nevertheless, emerging research has largely focused on unguided or exploratory use of AI tools and has raised concerns that learners may become overly reliant on AI-generated outputs, leading to superficial processing and diminished critical engagement. Recent studies (e.g., Çelik 2025; Li and Alias 2025; Nagori et al. 2025) therefore call for more research on careful instructional design and pedagogical guidance to ensure that Gen-AI enhances, rather than replaces, meaningful cognitive activity.

In response to this call, the present study examines the guided use of Gen-AI to support research article reading in an MA-TEFL course. Guided use in this study refers to the integration of explicit instruction in

research article move structure and AI prompting strategies prior to Gen-AI-assisted reading. By equipping students with foundational analytical reading strategies, the instructional component was intended to promote active engagement with research articles and reduce the likelihood of overreliance on AI-generated explanations. Specifically, this study investigates the effectiveness of two widely used Gen-AI tools, ChatGPT and NotebookLM, as scaffolding tools for Taiwanese EFL students. These tools are selected because they represent two distinct forms of AI support: dialogic interaction and document-grounded assistance, respectively. By examining their use within a guided instructional framework, the study seeks to explore how different forms of Gen-AI support influence EFL learners' comprehension of discipline-specific texts. In addition to examining reading comprehension outcomes, the study also explores students' perceptions of the usefulness of the instructional support and AI tools, as well as their interaction patterns during tool use. The study is designed to answer the following research questions:

RQ1: Does guided use of ChatGPT or NotebookLM better support reading comprehension of research articles in Applied Linguistics?

RQ2: To what extent do students perceive the usefulness of the pre-experimental instruction and each tool?

RQ3: What interaction patterns do students exhibit when using each tool?

2. LITERATURE REVIEW

The following sections discuss how Sweller's CLT (2011) and Vygotsky's ZPD (1978) inform EFL students' use of Gen-AI in learning activities. A review of Rhetorical Structure Analysis (RSA) and its effects on reading comprehension is then presented to explain the rationale for incorporating an instructional component into Gen-AI-assisted reading.

2.1 Cognitive Load Theory (CLT) Perspective on Gen-AI

CLT posits that learning is constrained by the limited capacity of working memory and distinguishes among different types of cognitive

load. In Sweller's (2011) revised framework, cognitive load is categorized into two types: intrinsic load and extraneous load. Within this framework, Gen-AI may serve as a tool that helps learners manage both forms of cognitive load during complex learning tasks. First, Gen-AI has the potential to reduce extraneous cognitive load, which arises from the way information is presented rather than from the content itself (Li and Alias, 2025). Extraneous load is high when readers encounter unfamiliar terminology, dense academic language, lengthy explanations, or difficulties accessing relevant information efficiently (Sweller 2011). By providing immediate explanations of technical terms, rephrasing complex passages into more accessible language, generating concise summaries, and answering learners' questions on demand, tools such as ChatGPT may help reduce unnecessary processing demands and enable learners to devote more cognitive resources to understanding disciplinary concepts (Kasneci et al. 2023). By comparison, NotebookLM may help readers process information-rich texts more efficiently by highlighting key ideas and generating source-grounded summaries (Alisoy 2025; Reyna 2025).

In addition, Gen-AI may help learners cope with intrinsic cognitive load, which stems from the inherent complexity of the learning material (Li and Alias, 2025). Unlike extraneous load, intrinsic load cannot be eliminated because it is determined by the nature of the content itself (Sweller 2011). Academic research articles, for example, often require readers to integrate multiple concepts, theoretical perspectives, methodological procedures, and empirical findings simultaneously, thereby imposing substantial intrinsic cognitive load. In this regard, Gen-AI may function as a form of cognitive scaffolding by segmenting information into smaller conceptual units, generating explanations tailored to learners' needs, highlighting relationships among key ideas, and creating visual representations such as concept maps.

Many of the proposed benefits of AI, particularly those associated with NotebookLM (Alisoy 2025; Reyna 2025), remain theoretically inferred from the tool's design features rather than directly supported by empirical evidence. Nevertheless, preliminary findings provide some support for the proposed role of Gen-AI in helping learners manage cognitive load (e.g., Çelik 2025; Nagori et al. 2025). For instance, Çelik (2025:26) integrated ChatGPT into a Task-Based Language Teaching (TBLT) framework,

termed “Intelligent-TBLT”, in which the tool functioned as a form of digital scaffolding. The study investigated 15 Turkish high school EFL learners engaged in a complex collaborative task involving a "deserted island" survival scenario. Participants worked in small groups to interpret advanced reading materials and collaboratively produce a 200-word survival plan under a strict 40-minute time limit. ChatGPT was introduced after an initial mid-task assessment to provide support. Using measures such as the NASA Task Load Index and reflective reports, the study found that learners initially experienced considerable difficulty due to the linguistic complexity of the materials, the cognitive demands of the task, and the time constraints imposed on task completion. Following the introduction of ChatGPT, participants reported lower levels of perceived cognitive load and greater confidence in completing the task. The findings suggest that AI-based scaffolding may help reduce extraneous cognitive load by providing immediate linguistic support and simplifying difficult texts, while also helping learners manage intrinsic cognitive load by organizing information, clarifying relationships among ideas, and supporting complex decision-making processes.

Despite its potential benefits, the use of Gen-AI in education also raises important concerns. Çelik (2025) cautions that learners may become overly reliant on AI-generated outputs, potentially leading to superficial processing of learning materials and reduced opportunities for independent reasoning and knowledge construction. Moreover, Li and Alias (2025) note that AI-generated responses may contain inaccuracies, fabricated information (i.e., hallucinations), or biased content, requiring learners to critically evaluate the reliability of Gen-AI outputs before accepting them as accurate. Other issues such as data privacy, academic integrity, and the ethical use of Gen-AI also warrant careful consideration in educational settings. From a CLT perspective, these concerns suggest that the cognitive resources made available through Gen-AI support should be redirected toward meaningful processing and knowledge construction rather than replaced by passive consumption of AI-generated content.

2.2 Vygotskian Constructivist Perspective on Gen-AI

According to Vygotsky's social constructivist theory (1978), knowledge is constructed through social interaction and guided support within the learner's ZPD, which is the gap between what learners can accomplish independently and what they can achieve with assistance from a more capable other. From this perspective, Gen-AI can be conceptualized as a form of "more capable other" that provides scaffolded support during the learning process. Empirical studies provide emerging evidence for this role. For example, Aljohani (2024) investigated how BA and MA students at the University of Jeddah used ChatGPT for EFL learning and perceived its usefulness. Participants reported using the tool for a variety of self-directed learning activities, including verifying answers, summarizing complex content, receiving language support, and improving writing. Notably, MA students relied on ChatGPT more frequently for advanced academic tasks such as writing and information synthesis, whereas BA students primarily used it for vocabulary development and foundational language skills. These findings suggest that learners at different academic levels may use Gen-AI as a form of self-directed scaffolding to support their language learning needs.

Additional evidence comes from Hong et al. (2025), who investigated the use of structured ChatGPT prompts to scaffold critical thinking in EFL writing among 92 second-year students enrolled in English writing courses at three vocational colleges in Thailand. The participants had intermediate English proficiency (approximately CEFR B1–B2) and engaged in an eight-week instructional intervention. Their findings indicate statistically significant improvements across multiple dimensions of critical thinking, including analysis, evaluation, and inference. Qualitative data further suggest that learners began to internalize these skills, transitioning from reliance on AI-generated prompts to more independent critical engagement. This progression reflects Vygotsky's concept of internalization, whereby externally provided support is gradually transformed into autonomous cognitive performance.

These studies suggest that Gen-AI can function as a mediational tool that supports learning within learners' ZPD. While Aljohani (2024) highlights learners' self-directed use of ChatGPT to support language learning, Hong et al. (2025) demonstrate that structured AI-mediated support can promote critical thinking and the gradual internalization of

higher-order cognitive skills. Methodologically, the two studies provide complementary forms of evidence. Aljohani relied on learners' self-reported experiences and perceptions, whereas Hong et al. employed an instructional intervention to examine learning outcomes over time. These findings provide preliminary support for the view that Gen-AI can facilitate learning through scaffolded assistance.

2.3 Move Analysis Instruction for Research Article Reading

To understand Move Analysis, it is useful to distinguish Rhetorical Structure Analysis (RSA) from the related concept of Text Structure Analysis (TSA). TSA is an approach to reading that focuses on identifying the logical organization of a text and the relationships among its ideas. It originates from the cognitive framework proposed by Meyer (1975) and Meyer and Poon (2001), in which reading comprehension is understood as the ability to recognize organizational patterns such as comparison–contrast, cause–effect, and problem–solution. Recognizing these patterns facilitates information recall and the construction of coherent mental representations. A substantial body of research has shown that explicit TSA instruction can enhance reading comprehension (e.g., Ghorbani Shemshadsara et al. 2019; Schwartz et al. 2017; Yeh et al. 2011).

RSA, by contrast, is a genre-based approach that extends beyond TSA's focus on logical organization to examine the communicative purposes of texts within disciplinary contexts. Grounded in Swales' (2004, 2014) genre theory, RSA focuses on identifying rhetorical "moves" and "steps," where a move represents a segment of text that serves a specific communicative function and a step realizes that function at a more detailed level. From this perspective, reading involves not only understanding what a text says, but also understanding how different sections contribute to the communicative goals of a discourse community.

Move Analysis represents a specific application of RSA that focuses on identifying the rhetorical moves and steps that characterize research articles. Rather than focusing solely on what a text says, Move Analysis examines what each section does and how it accomplishes its communicative purpose, such as establishing a research territory, identifying a research gap, describing procedures, reporting findings, or

discussing implications. Scholars have extensively examined the move structure of research articles in Applied Linguistics, providing valuable insights into the rhetorical organization of different article sections. The Create-A-Research-Space (CARS) model (Swales 2014) provides a widely used framework for analyzing research article introductions, consisting of three moves: establishing a territory, establishing a niche, and occupying the niche. These moves have been empirically validated in applied linguistics research (Khany and Tazik 2010). Extending rhetorical move analysis to other sections, Hu (2010) identified four key moves in literature reviews, with particular emphasis on surveying previous research and creating a research niche. Similarly, Wannaruk and Amnuai (2016) outlined four obligatory moves in the Methods section (e.g., participants, instruments, data collection procedures, data analysis), while Yang and Allison (2003) identified recurring patterns in Results and Discussion sections, including the cyclical use of reporting and commenting on results. The Conclusion section typically synthesizes the study by summarizing, evaluating, and drawing inferences. In sum, these studies suggest that research articles exhibit relatively predictable rhetorical patterns that can be made explicit through instruction.

Empirical evidence for the pedagogical value of genre-based rhetorical instruction is provided by Hyon (2001). Conducted in a 12-week EAP reading course, the study followed eight adult non-native English-speaking learners from diverse academic backgrounds and investigated their use of genre knowledge one year after instruction. The course introduced four genres (i.e., hard news articles, feature articles, textbooks, and research articles) through explicit instruction on content, rhetorical structure, language style, and communicative purpose, supported by activities such as text analysis and strategic reading practice. The findings suggest that learners developed durable “genre frames” that supported text interpretation, strategic reading, and transfer to academic writing. However, the study also noted potential limitations, including overgeneralization and the risk of overly prescriptive genre models.

Building on this evidence, Move Analysis instruction may serve as a useful form of guidance for Gen-AI-assisted reading by making the rhetorical organization of research articles explicit. Such guidance may encourage more strategic engagement with research articles and more

purposeful use of Gen-AI tools during reading. Research examining the integration of Move Analysis instruction with Gen-AI-assisted reading of research articles remains scarce. To address this gap, the present study incorporates Move Analysis instruction as a structured form of guidance prior to Gen-AI-assisted reading.

3. METHODS

This study employed a counterbalanced within-subject crossover design to examine the use of ChatGPT and NotebookLM in supporting EFL students' comprehension of Applied Linguistics research articles. The following sections describe the participants, pre-experimental instruction, reading materials, instruments, and data collection procedures.

3.1 Participants

This study involved 12 participants aged 21 to 25 years: six MA-TEFL graduate students and six senior undergraduate English majors enrolled in a 4+1 pre-master's program in Taiwan. None of the participants were native speakers of English, and all reported having no prior experience reading research articles. While all participants reported regular use of ChatGPT, only a few had prior experience with NotebookLM, making it a relatively unfamiliar tool.

3.2 Pre-Experimental Instruction

Before completing the reading tasks, all participants attended a 50-minute instructional session. During the session, they reviewed academic vocabulary commonly found in research articles and received training in Move Analysis and AI prompting strategies. The instruction was delivered by the researcher, who has many years of experience teaching graduate-level EFL reading courses.

Coxhead's (2000) Academic Word List (AWL), which comprises 570 word families commonly used in academic texts, was introduced as a foundation for the instruction. The focus was placed on reporting verbs

with nuanced meanings (e.g., *assert*, *argue*, *claim*, *state*, *hypothesize*, *indicate*, and *suggest*) to help participants distinguish varying degrees of certainty and rhetorical stance. In English for Academic Purposes (EAP), these verbs are often categorized as hedges (e.g., *suggest*, *indicate*, *hypothesize*), which convey tentativeness, and boosters (e.g., *assert*, *argue*, *claim*), which express stronger commitment.

A Move Analysis handout, developed from prior research on the rhetorical structure of research articles (e.g., Swales 2004, 2014; Hu 2010; Wannaruk and Amnuai 2016; Yang and Allison 2003), was used to introduce the move structure of different sections of empirical Applied Linguistics research articles. The session began with an overview of the macro-structure (Introduction, Literature Review, Methods, Results, and Discussion/Conclusion), followed by a systematic explanation of the moves and steps associated with each section. Guided discussions helped participants recognize relationships among moves. The instruction targeted three levels of rhetorical competence: (1) global awareness - understanding the overall structure of research articles; (2) functional awareness - identifying specific moves and steps (e.g., “Establishing a Niche,” “Indicating a Gap”); and (3) relational awareness - understanding how moves are logically connected (e.g., how identifying a gap leads to the statement of research purposes). The handout is provided in the Appendix.

Participants also received explicit guidance on prompting strategies tailored to the features of ChatGPT and NotebookLM and informed by their understanding of rhetorical structure. Example prompts focused on explaining unfamiliar terminology and concepts, summarizing article sections, identifying study purposes and research questions, explaining methodological procedures, interpreting findings, and clarifying the significance of results. They were further introduced to techniques for refining prompts through follow-up questions (e.g., identifying the paragraph that states the study purpose or explaining how the purpose is linked to a research gap) and for evaluating AI-generated responses against the original text (e.g., verifying whether reported findings were statistically significant by consulting the Results section). The goal was to help participants use AI strategically as a reading aid while drawing on

their understanding of rhetorical structure rather than relying on AI-generated responses as substitutes for reading the articles themselves.

3.3 Reading Materials

Two research articles on pragmatic transfer were selected for the experiment. The articles were comparable in length (approximately 8,000 words) and difficulty (Flesch Reading Ease ≈ 49), indicating college-level readability. Article A, *Exploring Request Strategies in Austrian Italian Learners: Pragmatic Transfer Insights* (Brocca and Nuzzo 2024), consisted of 7,983 words and 770 sentences (Flesch Reading Ease = 49.85) and was used in Experimental Session 1. Article B, *Disagreement by Chinese Speakers of English: Evidence of Pragmatic Transfer* (Zhu and Wang 2022), consisted of 8,089 words and 752 sentences (Flesch Reading Ease = 48.62) and was used in Experimental Session 2.

3.4 Instruments

Three instruments were used to collect data: reading comprehension tests, a perception-of-usefulness questionnaire, and on-screen recordings.

3.4.1 Reading comprehension tests

Two reading comprehension tests (RCTs) were developed for the respective articles to assess participants' comprehension. Each test consisted of 20 multiple-choice items targeting key content from different sections of the articles. The items were aligned with obligatory moves and selected steps identified through Move Analysis. A sample item is presented below.

- Q17. What major difference was found between L1 English and L2 English speakers? (Results – Move 2 - Presenting findings)
- A. L1 speakers used more words when expressing disagreement.
 - B. L2 speakers used more words when expressing disagreement.
 - C. L1 speakers used significantly stronger disagreement.
 - D. L2 speakers used significantly stronger disagreement. (Correct answer)

The two RCTs were designed in parallel to ensure comparability in format, length, and difficulty. Each correct response was awarded one point, yielding a maximum score of 20. Table 1 presents the alignment of test items with the corresponding moves and steps.

Table 1 Alignment of RCT Questions and Moves/Steps

#	Question Focus	Section	Move	Step
1	Research area	Introduction	Move 1: Establishing territory	Making topic generalization
2	Gap indicating	Introduction	Move 2: Establishing a niche	Indicating a gap in previous research
3	Study purpose	Introduction	Move 3: Occupying the niche	Announcing present research
4	Research questions	Introduction	Move 3: Occupying the niche	Announcing present research
5	Summary of previous research	Literature review	Move 2: Surveying previous research	Summarizing previous studies
6	Evaluation of previous research	Literature review	Move 2: Surveying previous research	Making evaluations
7	Limitation of prior studies	Literature review	Move 3: Creating a niche	Gap-indicating
8	Critique	Literature review	Move 3: Creating a niche	Counter-claiming
9	RQ function	Literature review	Move 4: Occupying the niche	Announcing aims

10	Nature of datasets	Methods	Move 1: Overview	Summarizing methods
11	Rationale for selection of participants	Methods	Move 3: Participants	Describing participants' backgrounds
12	Data collection site	Methods	Move 4: Research site	Describing location
13	Instruments	Methods	Move 5: Instruments	Specifying tools for data collection
14	Authenticity of data	Methods	Move 6: Data collection procedures	Explaining procedures
15	Statistical reasoning	Methods	Move 7: Data analysis	Explaining analytical purpose
16	Statistics	Results	Move 2: Reporting results	Presenting statistical results
17	Key finding 1	Results	Move 3: Comments on results	Interpreting results
18	Key finding 2	Results	Move 3: Comments on results	Interpreting results
19	Interpretation	Discussion	Move 3: Comments on results	Comparing results with literature
20	Evaluation of the findings	Discussion	Move 5: Evaluating the study	Highlighting significance

3.4.2 Perceptions-of-usefulness questionnaire

The perception-of-usefulness questionnaire was developed to examine participants' perceptions of the usefulness of the two Gen-AI tools and the

pre-experimental instruction for research article reading. The questionnaire consisted of 22 items rated on a 5-point Likert scale, including 10 items for ChatGPT, 10 for NotebookLM, and 2 for the pre-experimental instruction.

The 10 items for each tool assessed its perceived usefulness in helping participants: (1) identify the study purpose, (2) understand the significance of the study, (3) comprehend the main argument in the literature review, (4) clarify specialized terminology, (5) understand the methods, (6) interpret the results and findings, (7) reduce the time required to process difficult sections, (8) form a mental summary of the article, (9) enhance overall comprehension, and (10) increase confidence in understanding the text. The questionnaire was administered at the end of the experiment.

3.4.3 On-screen recordings

Behavioral data during the reading tasks were collected using Wondershare DemoCreator. At the beginning of the experiment, participants were instructed to download the application and were introduced to its key features. The software was used to record on-screen activities throughout each reading task, including time spent reading and interactions with ChatGPT and NotebookLM. Recording began when participants started reading the article abstract and continued until they completed their interactions with the assigned AI tool. Participants were given up to 50 minutes for each task, with a 10-minute break between the two sessions. As all 12 participants completed both tasks, a total of 24 screen recordings were collected.

The screen recordings and exported chat transcripts were subsequently examined by the researcher and coded according to the types of prompts and features used during interactions with ChatGPT and NotebookLM. To avoid inflating frequency counts, when more than one lexico-syntactic structure appeared within the same interaction turn on the same topic, only one instance was counted. When multiple clicks were made on a specific feature of NotebookLM, only one instance was recorded.

3.5 Data Collection Procedure

Following the 50-minute pre-experimental instruction on academic vocabulary and Move Analysis, participants completed two experimental sessions in which they read research articles using either ChatGPT or NotebookLM for up to 50 minutes. In Session 1, six participants read Article A (PDF format) using ChatGPT on a laptop, while the remaining six read the same article using NotebookLM. Their on-screen activities were recorded using Wondershare DemoCreator. After completing the reading activity, participants took the corresponding reading comprehension test (RCT). In Session 2, the procedure was repeated with Article B, with the conditions reversed: those who initially used ChatGPT switched to NotebookLM, and vice versa. At the conclusion of the experiment, all participants completed the perception-of-usefulness questionnaire.

4. RESULTS AND DISCUSSION

4.1 Reading Comprehension

Descriptive statistics were calculated for the RCT scores obtained with the two Gen-AI tools. Out of a maximum score of 20, NotebookLM yielded a higher mean score (mean = 16.75) and median score (median = 17.50) than ChatGPT (mean = 15.83, median = 16.00). NotebookLM also showed slightly lower variability (SD = 1.86) than ChatGPT (SD = 2.13). These results suggest that participants performed better and more consistently when using NotebookLM.

A Wilcoxon signed-rank test was conducted to examine whether there was a statistically significant difference in test scores between the two conditions. Because the sample size was small and the data were not assumed to be normally distributed, the nonparametric Wilcoxon test was deemed appropriate. After removing zero differences, 9 paired observations remained. The analysis showed that the sum of positive ranks (NotebookLM > ChatGPT) was $W^+ = 36$, whereas the sum of negative ranks (NotebookLM < ChatGPT) was $W^- = 9$. The test statistic was therefore based on the smaller of the two sums, $W = 9$. Using a normal approximation, the standardized test statistic was $Z = -1.60$, corresponding

to a two-tailed p-value of .11. This result indicates that higher scores were observed when using NotebookLM than when using ChatGPT, although the difference was not statistically significant at the conventional alpha level of .05. An effect size was subsequently calculated to assess the magnitude of the difference between the two conditions. Using the standardized statistic from the Wilcoxon signed-rank test, the effect size was computed as $r = Z/\sqrt{N}$, yielding a value of $r = .53$. This indicates a moderate-to-large effect size favoring NotebookLM. Despite the lack of statistical significance, the effect size suggests that the observed difference may be practically meaningful.

Table 2 presents the number and percentage of participants who correctly answered the questions under two conditions: using ChatGPT and using NotebookLM while reading the research articles. Several descriptive patterns emerge. First, the largest difference between the two conditions appears in Q8 (Critique), where 17% of participants answered correctly using ChatGPT, compared to 50% using NotebookLM, suggesting that NotebookLM may provide greater support for critique-related tasks. NotebookLM also tended to yield higher percentages of correct responses on other items requiring deeper understanding, including Q2 (Gap indicating, 75% vs. 58%), Q7 (Limitations of prior studies, 92% vs. 83%), Q15 (Statistical reasoning, 42% vs. 25%), and Q19 (Interpretation, 50% vs. 33%). Second, no observable differences were found for several items involving basic information retrieval. Performance was identical across both conditions, with accuracy reaching or approaching 100% for Q1 (Research area), Q3 (Study purpose), Q4 (Research questions), Q5 (Summary of previous research), Q9 (RQ function), Q10 (Nature of datasets), Q12 (Data collection site), Q13 (Instruments), Q14 (Authenticity of data), and Q16 (Statistics). This pattern suggests that both tools were similarly effective in supporting lower-level comprehension tasks involving explicit information retrieval. One exception was Q6 (Evaluation of previous research), for which only 67% of participants answered correctly in both conditions, indicating that this aspect of the article was comparatively more difficult to comprehend regardless of the AI tool used. Third, differences between the two conditions were generally small and inconsistent for mid-level comprehension tasks. For example, Q11 (Rationale for participants)

showed a slight advantage for NotebookLM (75% vs. 67%), whereas Q20 (Evaluation of findings) showed a slight advantage for ChatGPT (58% vs. 50%). Performances on Q17 (Key finding 1) and Q18 (Key finding 2) were also comparable across conditions. These results suggest that differences at this level of comprehension were minimal.

The individual item analysis above suggests a possible gradient in task difficulty and tool effectiveness. For higher-level tasks involving critique, gap indicating, limitations of prior studies, statistical reasoning, and interpretation, NotebookLM may show a tendency toward better performance. Both tools appear equally effective for lower-level tasks involving information retrieval, while differences in mid-level comprehension tasks are minimal. It should be noted that tasks requiring critique, statistical reasoning, and interpretation remained challenging for participants across both conditions.

Table 2 Number (%) of Participants Answering Correctly Under Two Conditions

#	Question focus	ChatGPT	NotebookLM
1	Research area	12 (100%)	12 (100%)
2	Gap indicating	7 (58%)	9 (75%)
3	Study purpose	12 (100%)	12 (100%)
4	Research questions	12 (100%)	12 (100%)
5	Summary of previous research	12 (100%)	12 (100%)
6	Evaluation of previous research	8 (67%)	8 (67%)
7	Limitation of prior studies	10 (83%)	11 (92%)
8	Critique	2 (17%)	6 (50%)
9	RQ function	12 (100%)	12 (100%)
10	Nature of datasets	12 (100%)	12 (100%)
11	Rationale for participants	8 (67%)	9 (75%)
12	Data collection site	12 (100%)	12 (100%)
13	Instruments	11 (92%)	11 (92%)
14	Authenticity of Data	12 (100%)	12 (100%)

15	Statistical Reasoning	3 (25%)	5 (42%)
16	Statistics	11 (92%)	11 (92%)
17	Key Finding 1	11 (92%)	12 (100%)
18	Key Finding 2	12 (100%)	11 (92%)
19	Interpretation	4 (33%)	6 (50%)
20	Evaluation of Findings	7 (58%)	6 (50%)

Overall, the findings suggest that NotebookLM may offer a modest advantage over ChatGPT in supporting reading comprehension, as indicated by its higher mean and median scores and a moderate-to-large effect size despite the absence of statistical significance. NotebookLM also showed a tendency to better support higher-order comprehension tasks, which require learners to move beyond surface-level processing. This pattern aligns with emerging theoretical discussions of NotebookLM as a tool designed to support document-grounded learning, synthesis, and knowledge organization (e.g., Alisoy 2025; Reyna 2025). These authors argue that NotebookLM’s ability to anchor responses in uploaded documents, generate structured summaries, and provide multimodal representations (e.g., concept maps, podcasts, and AI-generated questions) may make it particularly well suited to academic reading tasks that require the integration of information across a text. At the same time, several higher-order tasks remained challenging under both conditions, suggesting that effective engagement with research articles continues to require disciplinary knowledge and interpretive skills beyond those readily supported by Gen-AI tools. This interpretation is consistent with previous research emphasizing the importance of instructional guidance in supporting higher-order learning processes (Hong et al. 2025).

4.2 Perception of Usefulness

Descriptive statistics were calculated for the perception-of-usefulness questionnaire for the two Gen-AI tools. Overall, participants rated ChatGPT (mean = 3.97, standard deviation = 0.15, median = 4.00) and NotebookLM (mean = 3.92, standard deviation = 0.40, median = 3.85) similarly, suggesting comparable perceived usefulness. A Wilcoxon signed-rank test indicated that the difference in perceived usefulness

between the two tools was not statistically significant ($W = 30$, $Z = -0.71$, $p = .48$). The effect size was small ($r = .21$), suggesting that any difference in perceived usefulness was minimal.

Given that the difference in overall usefulness ratings between ChatGPT and NotebookLM was very small and not statistically significant, the item-level comparisons presented below should be interpreted as descriptive patterns rather than evidence of the superiority of either tool in specific areas. Table 3 presents the mean ratings of participants' perceived usefulness of ChatGPT and NotebookLM across ten reading-related functions. NotebookLM was rated higher in aspects related to structural and conceptual understanding, including identifying the study purpose (4.08 vs. 3.92), understanding the significance of the study (4.08 vs. 3.92), and making complex ideas easy to grasp (4.17 vs. 3.83). This suggests that NotebookLM may be particularly helpful in supporting higher-level organization and synthesis of information. In contrast, ChatGPT received higher ratings for items related to efficiency and clarity, such as clarifying specialized terminology (4.00 vs. 3.67), reducing the time required (4.42 vs. 4.00), understanding the methods (3.75 vs. 3.67), and interpreting the results and findings (4.33 vs. 4.17). These patterns suggest that ChatGPT may be perceived as useful for immediate comprehension and task efficiency. For several items, however, differences between the two tools were minimal or non-existent. For example, both tools received the same rating for comprehending the main argument in the literature review (3.58), and their ratings for enhancing overall comprehension (3.67 vs. 3.58) and increasing confidence (4.25 vs. 4.17) were very similar. These small differences reinforce the overall finding that participants viewed the two tools as broadly comparable in supporting research article reading. Taken together, the findings suggest a complementary relationship between the two tools: NotebookLM may better support global understanding and synthesis, whereas ChatGPT may facilitate clarification and efficiency. This distinction aligns with theoretical discussions of NotebookLM as a tool that supports structured, deeper comprehension (Reyna 2025), whereas ChatGPT aligns more closely with efficient, on-demand dialogic scaffolding (Aljohani 2024).

Table 3 Mean Ratings of Participants' Perceived Usefulness of Gen-AI Tools

#	Question Focus	ChatGPT	NotebookLM
1	Identify the study purpose	3.92	4.08
2	Understand the significance of the study	3.92	4.08
3	Comprehend the main argument in the literature review	3.58	3.58
4	Clarify specialized terminology	4.00	3.67
5	Understand the methods	3.75	3.67
6	Interpret the results/findings	4.33	4.17
7	Reduce the time required	4.42	4.00
8	Make complex ideas easy to grasp	3.83	4.17
9	Enhance overall comprehension	3.67	3.58
10	Increase confidence	4.25	4.17
Overall		3.97	3.92

With regard to the pre-experimental instruction, participants tended to rate both the academic English vocabulary instruction (mean = 3.83, SD = 0.39) and Move Analysis instruction (mean = 4.42, SD = 0.51) positively, with particularly higher ratings observed for the latter. This finding suggests that participants recognized the value of a rhetorical framework such as Move Analysis in supporting research article comprehension. This perception is consistent with Hyon's (2001) findings that students receiving genre-based reading instruction reported greater attention to rhetorical features and improved confidence in reading academic texts. The findings further extend this line of work by suggesting that Move Analysis instruction may provide a framework that supports learners' engagement with Gen-AI tools in more purposeful and academically relevant ways.

4.3 Observed Behaviors

All 12 participants' behavioral data during the AI-assisted reading tasks were analyzed. The recordings ranged from approximately 21 to 33

minutes in length, with an average duration of 27.6 minutes per task. Participants followed a relatively consistent sequence of actions across both tools: they first read the abstract (1–3 minutes), then scanned the research article (5–10 minutes), and subsequently uploaded the PDF to interact with the Gen-AI tool for approximately 15–20 minutes.

Despite this similar procedural pattern, analysis of the coded interaction data revealed distinct differences in how participants used the two Gen-AI tools. Table 4 presents the frequency counts of prompts and features used during participants' interactions with the tools. When using ChatGPT, participants primarily generated prompts to support comprehension. The most frequent prompt type was clarification of concepts or terms (25 instances), followed by requests for article summaries (12 instances). Prompts related to identifying the study purpose and research questions (8 instances) and gap-indicating (8 instances) were also common, while requests for method summaries were less frequent (5 instances). In contrast, when using NotebookLM, participants relied more on built-in features than on self-generated prompts. The most frequently used features were article summaries (12 instances) and concept maps (12 instances), followed by flashcards (10 instances), podcasts (9 instances), and AI-generated quiz questions (8 instances).

Table 4 Frequency Counts of Prompts or Features Used by Participants When Interacting with Gen-AI Tools (From High to Low)

Prompts Used When Interacting with ChatGPT	Frequency
Clarification (e.g., Pragmatic transfer, Z test, DCT)	25
Summary of the article	12
Study purpose research questions	8
Gap-Indicating	8
Summary of the methods	5
Features Used for Interacting with NotebookLM	Frequency
Summary of the article	12
Concept maps	12
Flashcards	10
Podcast	9
AI-generated quiz questions	8

The overall pattern suggests that although the use of both tools followed a similar reading workflow, learners engaged with them in qualitatively different ways. This finding is consistent with Hong et al.'s (2025) emphasis on scaffolding as a key component of AI-supported learning. While ChatGPT facilitated learner-initiated, responsive interactions through question generation and clarification, NotebookLM encouraged engagement through structured, multimodal features designed to support knowledge organization and synthesis (Alisoy 2025; Reyna 2025). These differences likely reflect the distinct affordances embedded in the design of the two tools.

5. CONCLUSION

This study investigated how explicit instruction in Move Analysis and AI prompting strategies, combined with the use of ChatGPT and NotebookLM, supported EFL students' comprehension of research articles. The findings suggest that NotebookLM may provide a modest advantage for certain higher-order comprehension tasks, whereas both tools were perceived as similarly useful overall. Participants also rated the instructional component positively, particularly the Move Analysis instruction. Finally, the two tools supported qualitatively different forms of engagement, with ChatGPT facilitating dialogic interaction and NotebookLM promoting structured, feature-based exploration of research articles.

The study contributes to the growing literature on AI-assisted academic reading by comparing ChatGPT and NotebookLM in an EFL context, examining both learning outcomes and interaction patterns, and highlighting the value of combining Gen-AI with explicit instructional support, particularly Move Analysis. These findings suggest that the key issue is not simply which tool is better, but rather how different forms of AI support can facilitate different aspects of academic reading within specific instructional contexts.

Several limitations should be acknowledged. A major limitation of this study is the absence of a non-AI control group. Consequently, the findings

allow only for a comparison between ChatGPT and NotebookLM and do not permit conclusions regarding the overall effectiveness of Gen-AI-assisted reading compared with traditional reading approaches. In addition, the small sample size limits the generalizability of the findings and reduces the statistical power. As a result, the findings should be interpreted as preliminary, and potentially meaningful differences between ChatGPT and NotebookLM may not have reached statistical significance due to the limited number of participants. The relatively short duration of the intervention represents another limitation. Because the activities were completed within a single instructional session, the findings capture only the immediate effects of Gen-AI-assisted reading on comprehension, perceptions, and interaction behaviors. The study therefore provides no evidence regarding the long-term impact of repeated use of ChatGPT or NotebookLM on academic reading development. Finally, the behavioral analysis focused on the frequency of prompts and feature use rather than the time participants spent engaging with each feature. As a result, it identified the prompts and features used most frequently but could not capture differences in the depth or duration of engagement.

Future research should address these limitations and extend the current findings in meaningful ways. First, the inclusion of a non-AI control group would help determine whether Gen-AI leads to measurable improvements in reading comprehension beyond traditional approaches. Second, larger-scale investigations with greater statistical power could clarify whether the observed advantage of NotebookLM represents a stable effect. Third, longitudinal designs could examine how sustained use of AI tools influences the development of reading strategies, comprehension, and academic literacy over time. Such studies could also explore how learners' AI interaction practices evolve with continued instruction and use. Finally, future research could move beyond frequency counts by incorporating time-on-task measures for specific prompts and AI features. This would provide a more detailed understanding of the depth and duration of learners' engagement and allow researchers to examine whether engagement with particular prompts or features is associated with reading comprehension performance.

On a broader scale, the present study points to the need for innovative educational practices. Although participants received guidance on prompting strategies, many relied on only brief prompts when using ChatGPT and depended heavily on NotebookLM's built-in features. Moreover, higher-order tasks remained challenging under both conditions. These patterns suggest that students may benefit from continued guidance that helps them engage with Gen-AI technologies in more strategic and purposeful ways. As Gen-AI becomes increasingly integrated into higher education, attention may need to shift from whether such tools should be used to how they can be aligned with instructional objectives and supported through effective pedagogical design.

REFERENCES

- Alisoy, Hasan. 2025. Can NotebookLM support English language learners? A theoretical perspective on AI tools in education. *Porta Universorum* 1:25-55. <https://doi.org/10.69760/portuni.0106003>
- Aljohani, Waad. 2024. ChatGPT for EFL: Usage and perceptions among BA and MA students. *Open Journal of Modern Linguistics*, 14:1119-1139. <https://doi.org/10.4236/ojml.2024.146059>
- Çelik, Ferdi. 2025. Intelligent-TBLT for EFL learners' reading and writing skills: A chaos complexity theory lens. *Computer-Assisted Language Learning Electronic Journal*, 26:23-45.
- Coxhead, Averil. 2000. A new academic word list. *TESOL Quarterly*, 34:213-238. <https://doi.org/10.2307/3587951>
- Ghorbani Shemshadsara, Zahra, Touran Ahour, and Nasrin Hadidi Tamjid. 2019. Raising text structure awareness: A strategy of improving EFL undergraduate students' reading comprehension ability. *Cogent Education*, 6, Article 1644704. <https://doi.org/10.1080/2331186X.2019.1644704>
- Hong, Hui, Poonsri Vate-U-Lan, and Chantana Viriyavejakul. 2025. Generative AI-mediated scaffolds for enhanced critical thinking in EFL writing. *Edelweiss Applied Science and Technology*, 9:43-54. <https://doi.org/10.55214/25768484.v9i6.7751>
- Hu, Jian. 2010. The schematic structure of literature review in research articles of applied linguistics. *Chinese Journal of Applied Linguistics*, 33:15-27.
- Hyon, Sunny. 2001. Long-term effects of genre-based instruction: A follow-up study of an EAP reading course. *English for Specific Purposes*, 20:417-438. [https://doi.org/10.1016/S0889-4906\(01\)00019-9](https://doi.org/10.1016/S0889-4906(01)00019-9)

- Kasneci, Enkelejda, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, Article 102274. <https://doi.org/10.1016/j.lindif.2023.102274>
- Khany, Reza, and Khalil Tazik. 2010. A comparative study of introduction and discussion sections of sub-disciplines of applied linguistics research articles. *Journal of Research in Applied Linguistics*, 1: 97-122.
- Li, Chenyu, and Amelia Alias. 2025. Generative artificial intelligence and cognitive load in second language learning: A narrative review. *International Journal of Academic Research in Progressive Education and Development*, 14:2459-2465. <https://doi.org/10.6007/IJARPED/v14-i4/27007>
- Meyer, Bonnie JF. 1975. *The organization of prose and its effects on memory*. Amsterdam; Oxford: North-Holland Publishing Company.
- Meyer, Bonnie JF., and Leonard W. Poon. 2001. Effects of structure strategy training and signaling on recall of text. *Journal of Educational Psychology*, 93:141-159.
- Nagori, Viral, Samir Mahmud, Shimi Zhou, and Lakshmi Iyer. 2025. Measuring cognitive load among students during Gen AI-enabled data analytics problem-solving. In *Proceedings of the 2025 Pre-ICIS SIGDSA Symposium*, Nashville, Tennessee, USA, pp. 1-7.
- Reyna, Jorge. 2025. The potential of Google NotebookLM for teaching and learning. In *Proceedings of the eLearn 2025 World Conference on E-Learning*, Bangkok, Thailand.
- Schwartz, Ana Isabel, Laura Mendoza, and Bonnie Meyer. 2017. The impact of text structure reading strategy instruction in a second language: Benefits across languages. *The Language Learning Journal*, 45:263-281. <https://doi.org/10.1080/09571736.2013.837092>
- Swales, John M. 2004. *Research Genres: Explorations and Applications*. Cambridge University Press.
- Swales, John M. 2014. Genre analysis: English in academic and research settings (originally Cambridge: Cambridge University Press, pp. 45-47, 52-60). In *The Discourse Studies Reader: Main currents in theory and analysis*, eds. Johannes Angermüller, Dominique Maingueneau, and Ruth Wodak, pp. 306-316. John Benjamins Publishing Company.
- Sweller, John. 2011. Cognitive load theory. In *Psychology of Learning and Motivation*, ed. B. H. Ross, Vol. 55, pp. 37-76. Academic Press. <https://doi.org/10.1016/B978-0-12-387691-1.00002-8>
- Vygotsky, Lev S. 1978. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.
- Wannaruk, Anchalee, and Wirada Amnuai. 2016. A comparison of rhetorical move structure of applied linguistics research articles published in international and national Thai journals. *RELC Journal*, 47:193-211.
- Yang, Ruiying, and Desmond Allison. 2003. Research articles in applied linguistics: Moving from results to conclusions. *English for Specific Purposes*, 22:365-385. [https://doi.org/10.1016/S0889-4906\(02\)00026-1](https://doi.org/10.1016/S0889-4906(02)00026-1)

Yeh, Li-Hao, Ana I. Schwartz, and Aaron L. Baule. 2011. The impact of text-structure strategy instruction on the text recall and eye-movement patterns of second language English readers. *Reading Psychology*, 32:495–519.
<https://doi.org/10.1080/02702711.2010.495621>

Chia-Hui Chiu
Department of Foreign Languages and Literature
Tunghai University
Taichung, Taiwan 407, ROC
chiahuichiu@thu.edu.tw

APPENDICES

Move Analysis of Research Article Structure in Applied Linguistics Part 1: Introduction Section Based on Swales' (2014) CARS Model

Move	Step
Move 1 – Establishing territory to commence research	✓ Claiming centrality ✓ Making topic generalization ✓ Reviewing the related studies
Move 2 – Establishing a niche	✓ Counter-claiming ✓ Indicating a gap ✓ Question raising
Move 3 – Occupying the niche	✓ Outlining purposes ✓ Announcing present research ✓ Announcing principal findings ✓ Indicating research article structure

Part 2: Literature Review Section Based on Hu's (2010) Framework

Move	Step
Move 1 – Establishing a thematic territory	✓ Making topic generalization ✓ Claiming centrality ✓ Giving background information

Move 2 – Surveying and summarizing previous research	✓ Constructing a reference to the published work ✓ Making positive and/or negative evaluations ✓ Making summary statements
Move 3 – Creating a research niche	✓ Counter-claiming ✓ Gap-indicating ✓ Question raising ✓ Asserting the relevancy ✓ Establishing or synthesizing a theoretical position or framework
Move 4 – Occupying the research niche	✓ Announcing the aims or research questions ✓ Announcing theoretical framework or positions ✓ Indicating research article structure ✓ Announcing the adoption or definitions of terminology ✓ Indicating possible findings ✓ Announcing research design or process

Part 3: Methods Section Based on Wannaruk and Amnuai (2016)

Move	Step
Move 1 – Overview	✓ Briefly summarize the research methods
Move 2 – Study purpose, research questions or hypotheses	✓ Explain the research objectives, the questions to be answered, and the hypotheses to be tested
Move 3 – Participants/Subjects	✓ Describe the participants and relevant background information
Move 4 – Research site	✓ Describe the location where the research was conducted
Move 5 – Instruments	✓ Specify the tools used for data collection
Move 6 – Data collection procedures	✓ Explain the data collection procedures
Move 7 – Data analysis	✓ Explain how the data were analyzed

Part 4: Results and Discussion Section - Yang and Allison (2003)

Move	Step
Move 1 – Provide preparatory information	✓ Provide relevant information to present the results
Move 2 – Report the results	✓ Present the results/findings using supporting evidence
Move 3 – Comment on the results	✓ Interpret the results
	✓ Compare the results with the literature
	✓ Evaluate the results
	✓ Explain possible reasons
Move 4 – Summarize the results	✓ Summarize the results
Move 5 – Evaluate the study	✓ Point out the limitations of the study
	✓ Highlight the significance or strengths of the study

Move 6 – Point out future research	✓ Suggest directions for further research
---------------------------------------	---

Part 5: Conclusion Section Based on Yang and Allison (2003)

Move	Step
Move 1 – Summarize findings	✓ Summarize the overall research findings
Move 2 – Evaluate the study	✓ Point out the limitations of the study
	✓ Highlight the significance or strengths of the study
	✓ Evaluate the research methods
Move 3 – Draw inferences from the research	✓ Suggest directions for further research
	✓ Present pedagogical (teaching) implications

引導式使用生成式人工智慧於研究論文閱讀：
臺灣英語教學碩士課程學生之個案研究

邱嘉慧
東海大學

協助 EFL 學生理解研究論文之教學法尚待探討。為填補此缺口，本個案研究結合篇章結構分析與生成式人工智慧提示語教學，探討其對臺灣碩士生學術閱讀理解之影響。十二名學生接受教學後，採平衡交叉設計閱讀兩篇論文，交替使用 ChatGPT（對話式）與 NotebookLM（依文件內容），並檢視閱讀理解表現、工具有用性感知及互動行為。結果顯示，NotebookLM 閱讀理解得分較高，ChatGPT 有用性感知較高，但均未達統計顯著；受試者認為篇章結構分析教學有助益。學生用 ChatGPT 多進行概念釐清提問，用 NotebookLM 則多依賴摘要與 Podcast 功能，前者較能提供結構化理解支持，後者則有助即時理解。

關鍵字：篇章結構分析、生成式人工智慧、研究論文閱讀理解

INPUT AND PERFUMING: AN INTERDISCIPLINARY DIALOGUE BETWEEN KRASHEN'S THEORY OF SECOND LANGUAGE ACQUISITION AND THE *YOGĀCĀRA* COGNITIVE MODEL^{*}

Wei-Wei Zhao
Quanzhou Normal University

ABSTRACT

This study explores potential conceptual alignments between Stephen Krashen's Monitor Theory and the *Yogācāra* cognitive model. Through conceptual comparison and textual analysis, it examines whether *Yogācāra* concepts may provide a complementary interpretive framework for several unresolved issues in Krashen's account of second language acquisition. The analysis focuses on the Acquisition-Learning, Monitor, Input, Natural Order, and Affective Filter hypotheses. While Krashen emphasizes subconscious acquisition through comprehensible input (*i+1*), the mechanisms through which input becomes enduring implicit linguistic competence remain only partially specified. This study therefore examines the *Yogācāra* doctrines of the eight consciousnesses (*aṣṭavijñāna*; 八識) and the perfuming-seed mechanism (*vāsanā-bīja*; 熏習-種子) as possible explanatory resources. It suggests that acquisition may be interpreted in relation to the storehouse consciousness (*ālaya-vijñāna*; 阿賴耶識), while learning and monitoring may correspond to functions of the mental consciousness (*mano-vijñāna*; 意識). Historical accounts of Yijing and Xuanzang are further examined as illustrative cases. The study argues that the *Yogācāra* cognitive model offers a complementary philosophical perspective for understanding language internalization and cognitive development in SLA.

^{*} We sincerely thank the two reviewers for their insightful comments and constructive suggestions, which greatly improved the quality of this manuscript. Any remaining errors are the sole responsibility of the author.

Keywords: Second language acquisition, Monitor Theory, *Yogācāra*, eight consciousnesses, perfuming (*vāsanā*), comprehensible input

1. INTRODUCTION

1.1 Research Background and Rationale

An examination of the development of second language acquisition (SLA) research shows that, since the 1970s, the field has shifted from behaviorist to cognitive approaches. Within this transition, Stephen Krashen's Monitor Theory has become one of the most influential frameworks. Krashen (1981) argues that language development depends primarily on subconscious acquisition rather than conscious learning, emphasizing comprehensible input as the principal condition for developing linguistic competence. Although the theory has had substantial pedagogical influence, many scholars have questioned its explanatory adequacy because several key mechanisms remain insufficiently specified. In particular, the process through which comprehensible input becomes stable internal competence remains a theoretical "black box."

Buddhist *Yogācāra* philosophy offers a systematic model of cognition that may provide an alternative interpretive perspective on this issue. Rather than treating *Yogācāra* solely as a metaphysical system, contemporary scholarship has increasingly understood it as a sophisticated analysis of experience, cognition, and consciousness (Kochumuttom 1982; Lusthaus 2002). Its doctrines of the eight consciousnesses (*aṣṭa-vijñāna*; 八識) and the storehouse consciousness (*ālaya-vijñāna*; 阿賴耶識), in particular, have been interpreted as explaining the continuity of unconscious cognitive processes and the accumulation of latent dispositions (Waldron 2003). These concepts therefore provide potentially valuable resources for examining how external linguistic experience may gradually become enduring cognitive competence.

1.2 Research Objectives and Core Questions

This study explores conceptual alignments and functional correspondences between Krashen's Monitor Theory and the

Yogācāra cognitive model through a comparative analysis of their theoretical frameworks. Its primary objective is to demonstrate that Krashen's five hypotheses may be reinterpreted through selected *Yogācāra* concepts, providing an alternative perspective on the cognitive mechanisms underlying second language acquisition.

To achieve this objective, the study addresses the following research questions:

(1) Primary Research Question

How may Krashen's input-acquisition mechanism be reinterpreted through the *Yogācāra* perfuming-seed mechanism (*vāsanā-bīja*; 熏習-種子)?

(2) Secondary Questions:

a. What functional correspondences may be identified between the Monitor and the discriminative function of the mental consciousness (*mano-vijñāna*; 意識)?

b. How may the Affective Filter be interpreted in relation to self-attachment (*ātma-grāha*; 我執) and the afflicted mental factors associated with the afflicted mental consciousness (*kliṣṭa-manas*; 末那識)?

c. How may the Natural Order Hypothesis be interpreted in relation to the conditions and temporal dynamics governing linguistic seed maturation?

1.3 Research Methods and Materials

This study adopts conceptual comparison and textual analysis to establish an interdisciplinary dialogue between second language acquisition theory and the *Yogācāra* cognitive model. Rather than reducing one framework to the other, it examines their structural correspondences and functional parallels as distinct interpretive systems.

The SLA component is based primarily on Krashen's foundational works (1981, 1982, 1985), supplemented by major critiques and developments proposed by Gregg (1984), Swain (1985), and Ellis (1994, 1997). The *Yogācāra* component draws principally on the *Chéng Wéishì Lùn* (*Vijñaptimātratāsiddhi-śāstra*; 《成唯識論》) and the *Thirty Verses on Consciousness-Only* (*Triṃśikā-vijñaptimātratā*; 《唯識三十頌》),

supplemented by modern studies of *Yogācāra* philosophy. These classical sources are supplemented by modern studies by Kochumuttom (1982), Lusthaus (2002), and Waldron (2003), which provide important discussions of consciousness, cognition, unconscious processes, and the storehouse consciousness.

Historical illustrations are drawn primarily from Yijing's *A Record of the Buddhist Religion as Practised in India and the Malay Archipelago* (*Nánhǎi Jīgūi Nèifǎ Zhuàn*) and related accounts of Buddhist pilgrimage and monastic education. These sources provide historical evidence for the discussion of language learning, linguistic cultivation, and the concept of perfuming.

1.4 Research Scope and Limitations

This study focuses on the theoretical dialogue between Krashen's Monitor Theory and the *Yogācāra* cognitive model, with particular attention to the cognitive mechanisms underlying language acquisition. The discussion is confined primarily to the Dharmapāla (護法) tradition of *Yogācāra* as transmitted by Xuanzang, reflecting both the scope of the inquiry and the availability of relevant textual sources. Accordingly, the study relies on conceptual analysis and historical textual evidence rather than large-scale empirical data from contemporary pedagogical settings.

2. THE INPUT VIEW AND COGNITIVE FRAMEWORK OF KRASHEN'S THEORY OF SECOND LANGUAGE ACQUISITION

Stephen Krashen's Monitor Theory consists of five interrelated hypotheses that together explain the cognitive mechanisms underlying second language acquisition. This chapter outlines these hypotheses, focusing on the acquisition–learning distinction, comprehensible input, affective factors, and major scholarly critiques as the theoretical foundation for the interdisciplinary comparison developed in subsequent chapters.

2.1 The Acquisition–Learning Distinction

One of the defining features of Krashen’s Monitor Theory is the distinction between acquisition and learning, formalized in the Acquisition–Learning Hypothesis. According to Krashen (1981), adult second language development proceeds through two independent systems.

Acquisition is a subconscious process analogous to first-language acquisition. Learners focus primarily on meaning rather than linguistic form and are typically unaware of the internalization process. The resulting implicit knowledge underlies fluent language use, with grammatical competence manifested as an intuitive sense of correctness rather than explicit rule knowledge.

Learning, by contrast, is a conscious process involving formal instruction, rule learning, memorization, and error correction. It produces explicit metalinguistic knowledge rather than the implicit competence required for spontaneous communication.

Building on this distinction, the Monitor Hypothesis proposes that consciously learned knowledge functions only as an editor of language produced by the acquired system rather than as its source. According to Krashen (1982), monitor use depends on three conditions: sufficient time, attention to linguistic form rather than meaning, and knowledge of the relevant grammatical rule.

The Natural Order Hypothesis further proposes that grammatical structures are acquired in a relatively predictable developmental sequence independent of instructional order. Studies of grammatical morphemes suggest that learners consistently acquire certain structures earlier than others, indicating that language development follows an internally constrained progression rather than a purely pedagogical sequence. The implications of this developmental pattern will be revisited in Chapter IV through comparison with the *Yogācāra* account of linguistic seed maturation.

2.2 Comprehensible Input and $i+1$

Within Krashen’s theoretical framework, the Input Hypothesis constitutes the principal mechanism through which environmental

exposure contributes to second language acquisition. It proposes that acquisition occurs when learners comprehend language slightly beyond their current level of competence ($i+1$).

Here, i represents the learner's current interlanguage level, while $+1$ denotes the next stage of linguistic development. According to Krashen (1985), the Input Hypothesis is characterized by two central principles. First, acquisition is driven by meaning-focused comprehension rather than explicit grammatical instruction. As learners understand messages through contextual and extralinguistic cues, accompanying linguistic structures are gradually internalized, making comprehensible input the primary driver of acquisition. Second, Krashen argues that abundant comprehensible input naturally exposes learners to $i+1$, reducing the need for artificially sequenced instructional syllabi.

In *The Power of Reading*, Krashen further extends this view through Free Voluntary Reading, arguing that extensive self-selected reading promotes vocabulary growth, grammatical development, and spelling more effectively than traditional rule-based instruction. This low-anxiety, meaning-oriented exposure is therefore presented as an effective pathway to advanced language proficiency. How such meaning-focused input becomes enduring cognitive competence will be examined in Chapter IV through the *Yogācāra* concept of perfuming.

2.3 The Affective Filter

The Affective Filter Hypothesis extends the Input Hypothesis by proposing that successful language acquisition depends not only on the availability of comprehensible input but also on the learner's affective state. According to Krashen (1982), affective variables influence whether comprehensible input is successfully processed by the acquisition system.

Krashen identifies three principal affective variables: motivation, self-confidence, and anxiety. Learners with high motivation and self-confidence but low anxiety are more likely to benefit from comprehensible input because their affective filter remains relatively low. Conversely, low motivation, diminished self-confidence, and elevated anxiety raise the affective filter, reducing the effectiveness of otherwise comprehensible input.

Unlike the Monitor Hypothesis, which concerns the conscious use of explicit knowledge, the Affective Filter Hypothesis addresses psychological conditions that facilitate or impede acquisition. Krashen argues that affective variables do not alter the quality of linguistic input itself but regulate learners' receptivity to that input.

Although the hypothesis has had considerable pedagogical influence, Krashen does not specify the cognitive mechanisms through which affective factors influence acquisition. This unresolved issue constitutes one of the principal theoretical gaps examined in the present study and provides the basis for the comparative analysis developed in Chapter IV.

2.4 Major Scholarly Critiques of Monitor Theory

Monitor Theory has exerted considerable influence on second language pedagogy because of its conceptual simplicity and practical applicability. Nevertheless, its theoretical assumptions and explanatory scope have also generated sustained scholarly debate. Three issues have been particularly influential in shaping subsequent developments in SLA research.

2.4.1 Definitional ambiguity and falsifiability

Gregg (1984) argues that Krashen's distinction between acquisition and learning lacks sufficiently precise psychological definitions, resulting in conceptual ambiguity. McLaughlin (1987) likewise questions the empirical accessibility of the subconscious processes proposed by the theory. With respect to the Input Hypothesis, Horner (1987) further argues that the boundary between i and $i+1$ cannot be operationally specified with sufficient precision, making the hypothesis difficult to test empirically and raising concerns about its falsifiability.

2.4.2 The neglect of output and interaction

Krashen's treatment of output as a consequence rather than a cause of acquisition has been challenged by Swain (1985, 1995), whose Output Hypothesis argues that language production promotes syntactic

processing and facilitates noticing linguistic gaps. Similarly, Long's (1996) Interaction Hypothesis emphasizes that acquisition develops through negotiated interaction between interlocutors, suggesting that comprehensible input alone cannot fully explain language development.

2.4.3 The interface between explicit and implicit knowledge

Krashen adopts a non-interface position, maintaining that explicitly learned knowledge cannot be transformed directly into acquired implicit competence. Subsequent discussions, however, have proposed possible interactions between explicit and implicit knowledge. Gu Qiyi (2005) and Zeng Yuping (2009) argue that empirical findings may support such interaction, while Wen Qiufang (2010) further suggests that explicit knowledge can, under appropriate conditions, facilitate the development of implicit competence.

Taken together, these critiques do not diminish the pedagogical significance of Monitor Theory. Rather, they identify theoretical questions concerning the cognitive mechanisms underlying language acquisition, thereby providing the point of departure for the comparative analysis developed in the following chapters.

3. THE COGNITIVE MODEL OF BUDDHIST *YOGĀCĀRA*: EIGHT CONSCIOUSNESSES, PERFUMING, AND SEEDS

The *Yogācāra* tradition provides a systematic model of cognition that offers valuable interpretive resources for examining cognitive processes in second language acquisition. Rather than serving as an alternative to contemporary SLA theory, the present study employs selected *Yogācāra* concepts as a complementary framework for exploring conceptual alignments and functional correspondences with Krashen's Monitor Theory.

This chapter introduces the eight-consciousness model together with the doctrines of perfuming, seeds (*bīja*; 種子), the afflicted mental consciousness, and the three natures (*tri-svabhāva*; 三自性). These concepts establish the theoretical foundation for the subsequent

comparative analysis by clarifying how *Yogācāra* conceptualizes the continuity, transformation, and conditioning of cognitive processes relevant to language learning.

3.1 The Eight Consciousnesses and the Analysis of Cognitive Structure

The *Yogācāra* cognitive model analyzes human cognition through a hierarchical structure of eight consciousnesses, each performing distinct cognitive functions. Unlike many contemporary psychological models that distinguish only between conscious and unconscious processes, *Yogācāra* presents a multi-layered cognitive architecture extending from sensory perception to the deepest repository of latent cognitive potential.

The first five sense-consciousnesses (*pañca-vijñānāni*; 前五識)—eye, ear, nose, tongue, and body-consciousness—are responsible for directly apprehending sensory information from the external environment. Within the present comparative framework, ear-consciousness and eye-consciousness may be understood as the primary channels through which phonological and orthographic input initially enters the cognitive system during second language acquisition.

The sixth consciousness, the mental consciousness, performs functions of analysis, conceptualization, and discrimination based on information received from the sensory consciousnesses. Rather than corresponding directly to any single construct in SLA theory, its cognitive functions exhibit meaningful functional correspondences with aspects of Krashen's learned system and Monitor, particularly in conscious analysis, rule processing, and discriminative cognition (*vikalpa*; 分別).

The deeper layers of the cognitive system consist of the afflicted mental consciousness and the storehouse consciousness. The afflicted mental consciousness continuously generates self-referential attachment by appropriating the storehouse consciousness as a permanent self. Within the present comparative framework, this function provides an interpretive perspective on affective processes in language learning, including anxiety, self-evaluation, and resistance to linguistic risk-taking. The storehouse consciousness serves as the repository of latent cognitive potential in which experiential traces are preserved through the process of perfuming.

As Yang Weizhong (2014) observes, its capacity to retain and accumulate linguistic seeds constitutes the theoretical basis for the continuity of cognition and learning.

From the perspective of the *Yogācāra* cognitive model, the eight consciousnesses together describe a dynamic cognitive system in which perception, conceptual processing, affective attachment, and latent cognitive continuity interact continuously. This functional organization provides the conceptual foundation for the subsequent comparison with Krashen's cognitive framework.

3.2 The Theory of Perfuming–Seed–Activation: An Inquiry into the Mechanism of Information Internalization

To examine how external linguistic information may become internal competence, this section analyzes the *Yogācāra* cycle of perfuming–seed–activation (*pravṛtti*; 現行). This cycle provides a useful interpretive framework for understanding the dynamic process of information internalization.

Within *Yogācāra* philosophy, a seed refers to a latent causal potential stored in the storehouse consciousness that can give rise to subsequent cognitive manifestations. Lin Guoliang (2007) notes that seeds possess several defining characteristics, including perishing in each instant (剎那滅), coexistence with the resultant (果俱有), and continuous transformation (恆隨轉), indicating that latent cognitive potentials are dynamic rather than static.

This process may be understood through three interrelated stages:

(1) Activation perfuming seeds (現行熏種子):

When learners encounter comprehensible input, the mental consciousness engages in corresponding cognitive processing. This activation leaves residual habitual tendencies within the storehouse consciousness, thereby forming new seeds or reinforcing existing ones. In the present comparative framework, this mechanism offers a possible account of how repeated linguistic exposure contributes to the accumulation of implicit linguistic potential. It therefore exhibits a functional correspondence with Krashen's notion of acquisition, while remaining conceptually distinct from the original SLA construct.

(2) Seeds giving rise to activation (種子生現行):

Linguistic seeds stored in the storehouse consciousness may, upon encountering appropriate conditions such as a communicative context, manifest as language production. Since this manifestation occurs largely without deliberate analytical intervention, it provides an interpretive parallel to Krashen's account of spontaneous fluent language use.

(3) Seeds successively arising from seeds (種子生種子):

Within the storehouse consciousness, seeds undergo continuous renewal and transformation. This mechanism helps explain how previously acquired linguistic resources may interact, reorganize, and reinforce one another over time, even when external input is limited.

From this perspective, language acquisition may be understood in relation to the dynamics of perfuming. As Master Jingjie (2020) explains, perfuming is a gradual process of accumulation in which small increments produce substantial results. Each instance of comprehensible input may be interpreted as an act of perfuming, reinforcing linguistic seeds through repeated exposure. Sustained exposure and repeated perfuming may therefore contribute to the stabilization of linguistic seeds, whereas limited or inconsistent exposure may result in weaker consolidation and greater susceptibility to attrition.

3.3 Analysis of Afflicted Mental Factors and Mechanisms of Cognitive Interference

Successful perfuming is not assumed to result automatically from environmental input. The *Yogācāra* cognitive model explains this variability through afflicted mental factors, which provide a useful interpretive perspective on the cognitive conditions that may facilitate or impede language internalization and exhibit functional correspondences with Krashen's Affective Filter.

According to the *Baifa Mingmen Lun* (*Treatise on the One Hundred Dharmas*; 《百法明門論》), afflictions are classified into root afflictions (根本煩惱) and secondary afflictions (隨煩惱). Within the context of second language acquisition, attachment (貪), aversion (瞋), delusion (癡), pride (慢), and doubt (疑) may be understood as mental factors that influence cognitive processing. For example, anxiety arising from the fear

of making mistakes may be interpreted as reflecting the activation of self-pride (*ātma-māna*; 我慢) and self-love (*ātma-sneha*; 我愛) within the afflicted mental consciousness.

From this perspective, afflicted mental factors temporarily defile consciousness and interfere with the effective processing of linguistic input. When self-attachment becomes excessively strong, the afflicted mental consciousness may disrupt the operation of the mental consciousness, thereby reducing the effectiveness of subsequent perfuming. Wen (2010) similarly discusses the moderating influence of affective factors on second language acquisition. Viewed through the *Yogācāra* cognitive model, an elevated affective filter may therefore be interpreted as reflecting the interference of afflicted mental factors with the process of effective perfuming.

Repeated cognitive interference may also reinforce maladaptive linguistic habits. When inaccurate linguistic forms are repeatedly produced, the resulting activations undergo repeated perfuming in the storehouse consciousness, gradually stabilizing as persistent linguistic seeds. This process offers a possible interpretive perspective on fossilization (Selinker 1972), suggesting that persistent non-target forms may result not only from insufficient input but also from the repeated reinforcement of previously established linguistic patterns under particular cognitive and affective conditions.

3.4 The Three Natures and the Construction of Linguistic Discrimination

The theory of the three natures provides a further perspective on how language shapes cognition and conceptual discrimination. Within the present comparative framework, it offers an interpretive basis for understanding learner interlanguage and developmental errors.

(1) The nature of false conceptualization (*parikalpita-svabhāva*; 遍計所執性)

The nature of false conceptualization refers to the mistaken attribution of inherent existence to phenomena through conceptual discrimination based on names and linguistic categories (名言分別). In second language learning, learners often transfer patterns from their first language to the

target language. Such transfer errors may be interpreted as manifestations of false conceptualization.

(2) The nature of dependent arising (*paratantra-svabhāva*; 依他起性)

The nature of dependent arising describes phenomena as arising through the interaction of multiple causes and conditions. Within the present framework, learner interlanguage may be viewed as exhibiting a conceptual alignment with this principle, representing a dynamic and transitional system shaped by external input, first-language influence, and ongoing cognitive processing.

(3) The nature of absolute reality (*pariniṣpanna-svabhāva*; 圓成實性)

The nature of absolute reality refers to the realization of the true nature of phenomena through an understanding of dependent arising. In the context of second language acquisition, it may be cautiously interpreted as an analogy for highly proficient and relatively effortless language performance rather than as a direct theoretical equivalent.

Within this triadic framework, linguistic discrimination is understood as a progressive cognitive process. The mental consciousness organizes experience through names and conceptual distinctions, while repeated perfuming gradually transforms conscious discrimination into the subconscious operation of linguistic seeds within the storehouse consciousness. From this perspective, increasing language proficiency reflects not the elimination of conscious learning but its gradual integration into more automatic patterns of cognitive processing through repeated perfuming.

3.5 Summary

This chapter has outlined the principal components of the *Yogācāra* cognitive model that are relevant to the present study. Through the analysis of the eight-consciousness framework, the mechanism of perfuming and seeds, afflicted mental factors, and the theory of the three natures, it has established a systematic account of cognitive continuity, internalization, and conceptual discrimination.

Rather than proposing a direct theoretical equivalence with contemporary second language acquisition theory, the present discussion has identified a set of *Yogācāra* concepts that provide a complementary

cognitive and interpretive framework for subsequent comparison with Krashen's Monitor Theory. These concepts furnish the theoretical foundation for examining how language input, cognitive processing, and internalization may be understood from an interdisciplinary perspective.

4. INPUT AND PERFUMING: AN INTERDISCIPLINARY DIALOGUE BETWEEN KRASHEN'S THEORY AND THE *YOGĀCĀRA* COGNITIVE MODEL

This chapter develops a systematic comparison between Stephen Krashen's Monitor Theory and the *Yogācāra* cognitive model. Building upon the theoretical foundations established in the preceding chapters, it examines the extent to which selected *Yogācāra* concepts may be interpreted as exhibiting conceptual alignments and functional correspondences with the principal constructs of Monitor Theory.

The comparative analysis focuses on four interrelated dimensions: input and perfuming, monitoring and discriminative cognition, affective filtering and self-attachment, and developmental sequencing and seed maturation. Through these comparisons, the chapter explores how the *Yogācāra* cognitive model may provide a complementary cognitive and interpretive perspective on the internalization of linguistic input while remaining conceptually distinct from contemporary second language acquisition theory.

4.1 Defining the Comparative Method and Analytical Framework

Before proceeding to the comparative analysis, it is necessary to clarify the methodological approach adopted in this study. The present study neither reduces the *Yogācāra* cognitive model to a precursor of modern psychology nor treats second language acquisition theory as a contemporary reformulation of *Yogācāra* philosophy. Instead, it employs a method of conceptual comparison and inter-reflection to identify structural correspondences between two distinct interpretive systems.

The analytical framework is organized around three dimensions: levels of consciousness, pathways of transformation, and mechanisms of

cognitive obstruction. As Jiang Jinlin (2010) observes, one of the central issues in second language acquisition research concerns the mechanisms through which linguistic knowledge becomes internalized. Within the present comparative framework, the *Yogācāra* account of perfuming provides an additional interpretive perspective for examining this process.

To facilitate a systematic comparison, Table 1 summarizes the principal conceptual alignments and functional correspondences between Krashen's Monitor Theory and the *Yogācāra* cognitive model. Rather than implying direct theoretical equivalence, the table serves as an analytical guide for the detailed comparisons developed in the following sections.

Table 1. Conceptual Alignments and Functional Correspondences Between Monitor Theory and the *Yogācāra* Cognitive Model.

Krashen's SLA Concept	<i>Yogācāra</i> Correspondence
Acquired System (Subconscious Process)	Activation of linguistic seeds (<i>bīja</i>) in the storehouse consciousness (<i>ālaya-vijñāna</i>)
Comprehensible Input	Perfuming through activation (<i>pravṛtti-vāsanā</i>)
Learned System & Monitor	Discriminative cognition (<i>vikalpa</i>) of the mental consciousness (<i>mano-vijñāna</i>)
Affective Filter	self-attachment (<i>ātma-grāha</i>) within the afflicted mental consciousness (<i>kliṣṭa-manas</i>)
Natural Order Hypothesis	Conditional seed maturation

4.2 Unconscious Acquisition and Linguistic Seeds in the Storehouse Consciousness

Although Krashen's distinction between the Acquired System and the Learned System captures an important pedagogical insight, the cognitive process through which comprehensible input becomes internal competence remains insufficiently specified. Within the present comparative framework, the Acquired System may be interpreted in

relation to the storehouse consciousness, where linguistic seeds are preserved as latent potentials of implicit linguistic competence.

This interpretation can be clarified through a cognitive sequence beginning with external linguistic input. Linguistic input, whether in the form of acoustic signals or written text, is first received through the sensory consciousnesses, especially ear-consciousness and eye-consciousness. At this initial level, input remains perceptual data rather than comprehended language.

The transition from external stimulus to comprehensible input occurs when the mental consciousness takes up these sensory impressions and performs semantic synthesis, contextual interpretation, and discriminative cognition. Through this process, the learner actively understands the meaning and relevance of the $i + 1$ input. Such comprehension is not merely a passive endpoint of exposure; it constitutes an activation, namely a current cognitive event within consciousness.

According to the *Yogācāra* cognitive model, every activation simultaneously contributes to perfuming the storehouse consciousness. The cognitive event of understanding input therefore functions as an active causal process through which linguistic experience leaves residual traces in the storehouse consciousness. This perfuming either forms linguistic seeds or reinforces previously established linguistic seeds corresponding to the linguistic structure being processed.

Once established within the storehouse consciousness, linguistic seeds undergo gradual maturation as relevant cognitive and communicative conditions develop. When a linguistic seed matures and encounters an appropriate communicative trigger, it may manifest as active language production without requiring explicit grammatical reflection or rule-governed monitoring. This process exhibits a functional correspondence with Krashen's account of spontaneous fluent production originating from the Acquired System.

Finally, the produced utterance may be processed again through the learner's own sensory consciousnesses, generating a recursive cycle of repeated perfuming. In this sense, subconscious acquisition may be interpreted as a continuous process in which comprehensible input activates cognitive processing, cognitive activation perfumes the

storehouse consciousness, linguistic seeds are formed and matured, and matured seeds subsequently manifest as fluent language use.

4.3 The Monitor and the Discriminative Grasping of the Mental Consciousness

Krashen defines learning as the conscious mastery of linguistic rules that functions as a monitor during language production. Within the present comparative framework, this monitoring function may be interpreted in relation to the discriminative function of the mental consciousness, rather than to the mental consciousness as a whole.

The mental consciousness possesses broad cognitive capacities, including semantic interpretation, inferential reasoning, memory retrieval, intentional reflection, and general linguistic processing. The conditions required for Monitor operation proposed by Krashen (1982)—sufficient time, focus on form, and knowledge of the rule—correspond to the circumstances under which the discriminative function of the mental consciousness can be deliberately engaged.

Accordingly, the Monitor should not be equated with the entirety of conscious cognition. Rather, it corresponds more specifically to the reflective, post-hoc, and rule-governed dimension of discriminative cognition, namely the deliberate application of explicit linguistic knowledge to monitor and revise language production. The Monitor therefore represents a specialized metalinguistic function rather than the totality of the mental consciousness.

Gregg (1984) criticized the Monitor Model for providing only a limited account of conscious cognitive operations. From the perspective of the *Yogācāra* cognitive model, this limitation may be understood in terms of the functional distinction between the mental consciousness and the storehouse consciousness. While explicit monitoring operates within conscious discriminative cognition, spontaneous language performance depends primarily on the activation of established linguistic seeds. Excessive reliance on monitoring therefore shifts attention toward conscious rule application rather than the natural manifestation of linguistic competence.

From this perspective, over-monitoring may be interpreted as the excessive engagement of discriminative cognition, which interrupts the spontaneous activation of linguistic seeds and reduces communicative fluency. This interpretation provides a complementary cognitive explanation for why excessive monitoring frequently results in hesitant and less natural language performance, particularly under conditions of increased cognitive load.

4.4 The Affective Filter and the Self-Attachment Interference of the Afflicted Mental Consciousness

Krashen's Affective Filter Hypothesis explains how affective variables influence the effectiveness of language acquisition but does not specify the underlying cognitive processes through which these variables exert their influence. Within the present comparative framework, these affective influences may be interpreted in relation to the self-attachment of the afflicted mental consciousness, which provides a complementary cognitive perspective on the operation of the Affective Filter.

It should be emphasized that the Affective Filter is not identical to the afflicted mental consciousness. Rather, the relationship between them is one of structural correspondence rather than theoretical equivalence. Whereas the Affective Filter describes the observable influence of affective variables on language acquisition, the afflicted mental consciousness refers to the deeper cognitive structure through which continual self-attachment generates emotional and cognitive interference.

(1) Self-attachment as the cognitive origin of affective interference

Within the *Yogācāra* cognitive model, the afflicted mental consciousness persistently appropriates the storehouse consciousness as an enduring self. This continual self-attachment gives rise to emotional reactions such as communication anxiety, fear of negative evaluation, excessive self-consciousness, and defensive self-protection. When learners perceive that their linguistic self-image is threatened, self-attachment intensifies, increasing the likelihood that affective factors will interfere with cognitive processing during language learning.

(2) The causal mechanism of dual-level cognitive interference

From the perspective of the *Yogācāra* cognitive model, affective interference operates as a continuous causal process rather than as two independent levels of disruption.

The first stage of interference occurs within the mental consciousness. Once self-attachment generates anxiety or defensive emotions, attentional resources become increasingly occupied by self-referential concerns. As a result, semantic interpretation, contextual integration, and discriminative cognition are disrupted, reducing the learner's capacity to process comprehensible input effectively.

The second stage follows directly from the first. Because effective perfuming depends upon successful cognitive processing within the mental consciousness, diminished comprehension weakens the subsequent perfuming of the storehouse consciousness. Linguistic experience therefore leaves a less stable cognitive imprint, resulting in weaker formation and reinforcement of linguistic seeds. Over time, this cumulative reduction in effective perfuming may limit the maturation of linguistic seeds and ultimately diminish the spontaneous manifestation of fluent language performance.

Accordingly, affective interference should be understood as a dynamic causal chain extending from self-attachment to conscious cognitive disruption, from weakened cognitive processing to reduced perfuming, and finally from insufficient seed reinforcement to slower development of implicit linguistic competence. Whereas Krashen primarily describes the pedagogical consequences of the Affective Filter, the *Yogācāra* cognitive model provides a complementary interpretation of the cognitive mechanism through which affective factors may influence both immediate language performance and the long-term internalization of linguistic input.

4.5 Natural Order and the Conditions and Timing of Linguistic Seed Maturation

Krashen's Natural Order Hypothesis proposes that language structures are acquired in a relatively predictable developmental sequence. Within the present comparative framework, this developmental sequence may be interpreted in relation to the gradual maturation of linguistic seeds under appropriate cognitive conditions and temporal development. Rather than

challenging the empirical findings of the Natural Order Hypothesis, the *Yogācāra* cognitive model offers a complementary interpretation of the cognitive processes through which developmental sequencing may emerge.

(1) Inherent temporality and seed maturation

Research on the Natural Order Hypothesis demonstrates that learners consistently acquire particular grammatical structures earlier than others, suggesting an internal developmental sequence independent of instructional order. Within the *Yogācāra* cognitive model, this sequence may be interpreted as reflecting differences in the maturation of linguistic seeds. Linguistic seeds stored in the storehouse consciousness mature gradually according to both cognitive conditions and temporal development rather than emerging simultaneously. The spontaneous appearance of a particular linguistic structure therefore indicates that its corresponding seed has reached a sufficient level of maturation for manifestation.

Seed maturation depends upon the interaction of external linguistic input and internal cognitive readiness. External input provides the conditions for continued perfuming, whereas the manifestation of a linguistic seed requires that sufficient prior perfuming has accumulated within the storehouse consciousness. Consequently, differences in developmental sequence may be understood as reflecting differences in the maturation of linguistic seeds rather than differences in the quantity of input alone. This interpretation is consistent with the *Yogācāra* cycle in which manifestations perfume seeds, and matured seeds subsequently give rise to further manifestations.

(2) Developmental readiness and condition dynamics

The SLA concept of developmental readiness may be interpreted through the *Yogācāra* notion of the immediately antecedent condition (等無間緣). From this perspective, each stage of cognitive development depends upon the successful completion of the preceding stage. Higher-level linguistic structures cannot become cognitively available until lower-level linguistic seeds have first manifested, stabilized, and undergone further perfuming within the storehouse consciousness.

This interpretation suggests that instructional input cannot bypass the learner's current developmental state. When linguistic input exceeds the

learner's cognitive readiness, the internal conditions necessary for effective perfuming have not yet been established. Developmental progression should therefore be understood as a condition-dependent process in which earlier linguistic manifestations provide the cognitive basis for the maturation of more advanced linguistic potentials.

(3) The natural syllabus and effortless unfolding

Krashen argues that instruction should respect the natural order of acquisition rather than impose an artificial grammatical sequence. This pedagogical position exhibits a functional correspondence with the *Yogācāra* view that cognitive transformation unfolds naturally when appropriate causes and conditions are present. If external input is incompatible with the learner's current stage of seed maturation, effective perfuming is unlikely to occur.

From this perspective, the Natural Order reflects the gradual unfolding of linguistic potential through the maturation of linguistic seeds under appropriate cognitive and temporal conditions. The pedagogical implication is not the abandonment of instruction but the cultivation of learning environments in which comprehensible input, cognitive readiness, and repeated perfuming converge to support the natural development of implicit linguistic competence.

4.6 Interlanguage and a Reinterpretation through the Three Natures

Within the present comparative framework, the theory of the three natures provides an alternative interpretation of learner interlanguage and learner errors.

From the perspective of the nature of dependent arising, learner interlanguage is not a collection of errors but a dynamic system arising from multiple conditions, including comprehensible input, linguistic seeds, first-language influence, learning motivation, and other cognitive factors. From this perspective, the errors exhibited at a particular stage represent the natural manifestation of linguistic seeds under the conditions prevailing at that stage.

The nature of false conceptualization, by contrast, arises when teachers or learners impose essentialized interpretations on these temporary developmental systems. For example, regarding learner errors

as evidence of deficient competence rather than as part of an evolving developmental process reflects the nature of false conceptualization. As Ellis (1994) argues, interlanguage is characterized by both systematicity and variability.

In contrast, the nature of absolute reality may be understood as an analogy for the stage at which learners have substantially overcome first-language interference and achieved relatively effortless and spontaneous language use through the maturation of linguistic seeds. Collectively, the three natures reinterpret interlanguage as a dynamic process of cognitive development rather than a linear progression from error to correctness.

4.7 Systematic Evaluation of the Central Research Questions

To ensure alignment between the research questions proposed in Chapter I and the comparative analyses developed in this chapter, this section systematically evaluates the primary and secondary research questions. Drawing upon the findings presented in Sections 4.2–4.6, the principal correspondences between Krashen's Monitor Theory and the *Yogācāra* cognitive model are summarized as follows.

4.7.1 Evaluation of the primary question: the input-acquisition mechanism through perfuming-seed model

The primary research question examines how the subconscious input–acquisition mechanism may be reinterpreted through the *Yogācāra* model of perfuming and linguistic seeds. Section 4.2 demonstrates that comprehensible input may be understood as a continuous cognitive process involving activation, perfuming, the formation and maturation of linguistic seeds, and their subsequent manifestation. Accordingly, subconscious acquisition may be interpreted as the gradual internalization of linguistic experience through the continuous development of linguistic seeds within the storehouse consciousness.

4.7.2 Evaluation of the first secondary question: the monitor and mental consciousness

Section 4.3 shows that the Monitor corresponds specifically to the discriminative function of the mental consciousness rather than to conscious cognition as a whole. This interpretation explains why excessive monitoring diverts attentional resources toward explicit rule application and thereby reduces the spontaneous manifestation of linguistic seeds.

4.7.3 Evaluation of the second secondary question: the affective filter and afflicted mental consciousness

Section 4.4 suggests that the Affective Filter may be interpreted through self-attachment within the afflicted mental consciousness. Affective interference operates as a causal process in which self-attachment disrupts cognitive processing, weakens subsequent perceiving, and ultimately limits the formation, maturation, and manifestation of linguistic seeds.

4.7.4 Evaluation of the third secondary question: natural order and seed maturation chronologies

Section 4.5 interprets the Natural Order Hypothesis as a condition-dependent developmental sequence governed by the gradual maturation of linguistic seeds. Linguistic development therefore depends not only on comprehensible input but also on the learner's cognitive readiness and the progressive maturation of linguistic seeds.

4.7.5 Evaluation of the fourth secondary question: interlanguage and the three natures

Section 4.6 reinterprets learner interlanguage through the theory of the three natures. Learner language is understood as a dynamic cognitive system arising from multiple conditions, while developmental errors reflect temporary manifestations of linguistic seeds rather than fixed deficiencies. This perspective reframes interlanguage as an evolving process of cognitive development.

4.8 Theoretical Implications and Transition to Historical Analysis

The foregoing analysis suggests that the *Yogācāra* cognitive model does not replace Krashen's SLA framework but provides a complementary cognitive explanation of the micro-processes underlying language acquisition. Whereas Krashen primarily emphasizes the environmental conditions facilitating acquisition, the *Yogācāra* cognitive model offers a more detailed account of how linguistic input is transformed into enduring cognitive potential.

The implications of this interdisciplinary comparison, however, cannot be fully evaluated through conceptual analysis alone. Historical and pedagogical evidence is required to examine whether the cognitive mechanisms proposed in this study can be identified in actual language-learning practices.

The following chapter therefore examines the experiences of Buddhist pilgrim-monastics, particularly Xuanzang and Yijing, to explore the historical and pedagogical implications of the input-perfuming model.

5. THE BUDDHIST MODEL OF LANGUAGE LEARNING AND ITS DIALOGUE WITH MODERN PEDAGOGY

Building upon the comparative framework established in Chapter IV, this chapter examines whether the proposed cognitive mechanisms may be identified in historical language-learning practices. It therefore turns to the experiences of Buddhist pilgrim-monastics during the Sui and Tang periods, particularly Xuanzang and Yijing, whose records provide rare and detailed accounts of intensive foreign-language learning in premodern Asia.

By analyzing their approaches to linguistic immersion, phonological training, linguistics (*śabdavidyā*; 聲明), and religious motivation, this chapter explores how the cognitive processes proposed in the preceding chapters may be interpreted within concrete historical contexts.

Drawing upon these historical cases, the chapter further considers the pedagogical implications of the Buddhist model of language learning for contemporary second language education. Particular attention is given to

the relationship among extensive comprehensible input, repeated perfuming, embodied recitation, affective regulation, and long-term motivational orientation. The historical analysis is intended to provide contextual illustrations of the proposed theoretical framework rather than empirical verification of its validity.

5.1 The Buddhist Model of Language Learning: Four Core Dimensions

The Buddhist model of language learning comprises four interrelated dimensions: perfuming, linguistics, correct pronunciation, and Dharma propagation. The language-learning practices of Buddhist pilgrim-monastics during the Sui and Tang dynasties, who traveled to India, Śrīvijaya (室利佛逝), and other regions, reveal a systematic learning pathway that may be interpreted through the cognitive framework developed in Chapter IV. Rather than relying primarily on explicit rule learning, this model integrates extensive comprehensible input (perfuming), systematic linguistic study, correct pronunciation, and the long-term motivation associated with Dharma propagation into a coherent process of language learning.

5.1.1 Situational acquisition and the perfuming-oriented view of learning

Yijing, in his works *A Record of the Buddhist Religion as Practised in India and the Malay Archipelago* (Nánhǎi jìguī nèifǎ chuán) and *Biographies of Eminent Monks of the Great Tang Seeking Scriptures in the Western Regions* (Dài Táng xīyù qiúfǎ gāosēng chuàn), documented the language preparation undertaken by pilgrim-monastics before formal scriptural study. Many monastics resided for several years in places such as Śrīvijaya (present-day Sumatra) or East India, attaining functional communicative fluency through sustained participation in daily chanting, listening to discourses, and everyday interaction within the monastic community.

This learning pathway exhibits a functional correspondence with the perfuming-oriented mechanism proposed in Section 4.2. Rather than beginning with explicit grammatical instruction, the monastics first

acquired linguistic competence through sustained auditory and visual exposure in meaningful communicative contexts. Through repeated comprehension and repeated perfuming, linguistic seeds were gradually formed, reinforced, and matured, ultimately supporting fluent language use with minimal reliance on conscious monitoring.

5.1.2 Linguistics, systematic knowledge, and the monitor

However, the linguistic competence of Buddhist monastics was not confined to oral communication. Their ability to translate Buddhist scriptures depended upon rigorous training in linguistics. During his studies at Nālandā Monastery, Xuanzang trained under Śīlabhadra and studied traditional Indian linguistic works such as the *Kāśikāvṛttī* (聲明記論).

Within the proposed comparative framework, this linguistic training exhibits a functional correspondence with the Monitor discussed in Section 4.3. Rather than replacing the acquisition process, systematic linguistic knowledge provided analytical resources for the mental consciousness during translation and doctrinal debate. Accordingly, the historical evidence suggests that extensive perfuming provided the foundation for language acquisition, whereas systematic grammatical study functioned as an auxiliary resource for conscious monitoring.

5.1.3 Correct pronunciation, oral transmission, and sound-centered learning

A distinctive feature of Buddhist pedagogy is its emphasis on sound and oral memory. From the syllabic practice of Siddham script to the oral transmission of mantras and incantations (*dhāraṇī*; 陀羅尼), language learning was closely associated with repeated recitation and embodied practice. This sound-centered approach may be interpreted as facilitating the continuous accumulation of linguistic seeds through repeated auditory exposure.

Unlike many modern approaches that emphasize textual analysis, the Buddhist model places greater emphasis on repeated recitation as a form of repeated perfuming. Rather than relying primarily on explicit linguistic

analysis, this embodied approach reinforces language learning through sustained auditory input and oral practice. Within the proposed comparative framework, this historical practice illustrates the perfuming-oriented learning mechanism discussed in Section 4.2.

5.1.4 Motivation, dharma propagation, and the affective filter

Krashen proposes that the affective filter influences the effectiveness of input internalization. For Buddhist pilgrim-monastics, the motivation to study Sanskrit was rooted in the aspiration to “continue the Buddha’s wisdom-life” and to “benefit sentient beings.” This enduring religious commitment may be understood as fostering a high level of integrative motivation.

Within the proposed comparative framework, the aspiration to propagate the Dharma may be interpreted as reducing self-attachment within the afflicted mental consciousness. As a result, monastics were better able to maintain receptivity to linguistic input despite hardship, failure, or ridicule. In this sense, a stable long-term purpose may help reduce affective interference, thereby supporting both effective perfuming and the continued maturation of linguistic seeds. This historical pattern provides a complementary perspective on how some pilgrim-monastics attained high levels of linguistic proficiency despite the challenges of cross-cultural learning.

It should be emphasized that the present discussion does not claim that the historical practices of Buddhist pilgrim-monastics were consciously based on the *Yogācāra* cognitive model reconstructed in this study. Rather, these historical cases serve as heuristic illustrations through which the structural and functional correspondences proposed in Chapter IV may be examined within concrete historical contexts. The present analysis should therefore be understood as an interpretive application of the theoretical framework rather than as a historical demonstration of causal influence.

5.1.5 Summary of the four dimensions

In summary, the Buddhist model of language learning may be characterized by four interrelated dimensions:

- (1) Perfuming: sustained, contextualized comprehensible input that supports the formation and maturation of linguistic seeds.
- (2) Linguistics (*śabdavidyā*): systematic linguistic knowledge serving as an auxiliary resource for conscious monitoring.
- (3) Correct pronunciation: repeated recitation and oral practice that reinforce repeated perfuming.
- (4) Dharma propagation: long-term religious aspiration that reduces affective interference and supports sustained language learning.

5.2 Implications for Contemporary Second Language Pedagogy from the Buddhist Model

The historical practices discussed above, interpreted through the proposed comparative framework, offer several pedagogical implications for contemporary second language education, including Chinese language teaching.

5.2.1 Curriculum design centered on perfuming-oriented input

Modern language instruction is often constrained by limited class time and therefore emphasizes short-term rule learning. From the perspective developed in this study, however, language competence depends primarily on repeated perfuming. Educators should create sustained environments for comprehensible input, such as free voluntary reading and immersive listening. Rather than emphasizing immediate grammatical correction, instruction should prioritize the long-term formation and maturation of linguistic seeds.

5.2.2 Positioning explicit rules as metacognitive refinement tools

This study does not advocate abandoning grammar instruction. Rather, grammatical knowledge should function as a metacognitive resource for conscious monitoring. As Krashen argues, rules should support monitoring rather than initiate speech. Accordingly, rule-first instruction should be avoided, and explicit knowledge should be introduced after learners have accumulated sufficient perfuming-based intuition.

5.2.3 Revitalizing sound, recitation, and embodied learning

Greater emphasis should be placed on sound, recitation, and oral practice. Through rhythmic recitation and vocal imitation, learners may reduce excessive reliance on conscious analysis while reinforcing repeated perfuming through sustained auditory and articulatory practice. Establishing a sound-based representation before full semantic analysis may further facilitate subsequent comprehension.

5.2.4 Affect, motivation, and long-term learning goals

Learning anxiety often arises from performance-oriented goals, such as examinations. Drawing inspiration from the motivation of Dharma propagation, educators should encourage learners to establish long-term purposes that extend beyond self-performance, such as cross-cultural communication or sustained engagement with the target language and culture. As learners shift their attention from self-performance to intrinsic purpose, affective interference may be reduced, thereby creating more favorable conditions for language acquisition.

5.3 Summary

This chapter has interpreted the language-learning practices of Buddhist pilgrim-monastics through the comparative framework developed in the preceding chapters, providing historical illustrations of the proposed structural and functional correspondences. The cases of Yijing and Xuanzang suggest that the Buddhist model of language learning offers a meaningful historical perspective on the input–perfuming framework and its pedagogical implications for contemporary second language education.

The present historical analysis is intended to complement, rather than empirically verify, the theoretical framework developed in this study. Its contribution lies in demonstrating how the proposed cognitive model may be interpreted within concrete historical

learning practices while offering pedagogical insights for contemporary language teaching.

6. DISCUSSION

This study has developed an interpretive framework based on the conceptual alignments and functional correspondences between Stephen Krashen's Monitor Theory and the *Yogācāra* cognitive model. Building upon the analyses presented in the preceding chapters, this chapter synthesizes the principal findings, discusses their theoretical and pedagogical implications, and considers the study's limitations and future directions.

6.1 Research Conclusions: Toward an Integrated Interpretation of Input and Perfuming

The analyses presented in this study suggest that Krashen's hypotheses on second language acquisition exhibit meaningful conceptual alignments and functional correspondences with selected principles of cognition articulated in the *Yogācāra* cognitive model.

(1) Correspondences in the structure of consciousness

The distinction between acquisition and learning may be interpreted through the complementary functions of the storehouse consciousness and the mental consciousness. Within this framework, spontaneous language use reflects the manifestation of linguistic seeds, whereas the Monitor represents conscious monitoring based on explicit linguistic knowledge.

(2) Interpretive correspondence in the mechanism of transformation

This study proposes an interpretation of comprehensible input through the mechanism of perfuming. Comprehensible input may be understood as contributing to the formation and maturation of linguistic seeds, thereby providing a complementary explanation of how external linguistic experience becomes enduring implicit linguistic competence.

(3) Ontological roots of the affective filter

The Affective Filter may be interpreted through the influence of self-attachment within the afflicted mental consciousness. From this

perspective, affective interference influences the effectiveness of perfuming and consequently the formation, maturation, and manifestation of linguistic seeds.

6.2 Theoretical and Practical Implications: The Interpretive Value of East Asian Philosophy of Consciousness

The principal theoretical contribution of this study lies in proposing an interpretive framework for second language acquisition rooted in the East Asian intellectual tradition. Rather than replacing existing SLA theories, the *Yogācāra* cognitive model provides a complementary perspective on the cognitive mechanisms underlying language acquisition, particularly the processes through which comprehensible input is internalized as enduring implicit linguistic competence.

From a pedagogical perspective, this study highlights the importance of integrating cognitive and affective dimensions in language teaching. Building upon the historical and theoretical analyses presented in the preceding chapters, it suggests that sustained comprehensible input, repeated perfuming, sound-based practice, and long-term learning purposes together provide conditions conducive to language acquisition. These implications are intended to complement, rather than replace, contemporary approaches to second language pedagogy.

6.3 Research Limitations and Future Directions

This study has several limitations. First, it is primarily theoretical, combining conceptual comparison with historical textual analysis. Although supported by historical cases such as Yijing and Xuanzang, it does not include large-scale empirical evidence from contemporary language-learning contexts. Second, the *Yogācāra* cognitive model encompasses a broad and complex intellectual tradition. While this study follows the Dharmapāla lineage transmitted by Xuanzang, space limitations have precluded a comprehensive discussion of the doctrinal differences among *Yogācāra* subschools, such as debates concerning the origin and existence of linguistic seeds.

Future research may proceed in three directions:

(1) Empirical Pedagogical Experiments:

Design experimental and control groups that compare instructional approaches integrating affective regulation with approaches emphasizing extensive comprehensible input alone, in order to examine their effects on language acquisition and the development of implicit linguistic competence.

(2) Qualitative Interview Studies

Conduct in-depth interviews with Buddhist monastic education systems and highly proficient second language learners to explore the roles of repeated perfuming and conscious monitoring in language learning.

(3) Development of Teaching Materials

Develop second language teaching materials informed by the theory of perfuming that emphasize sound, rhythm, and contextualized language internalization.

6.4 Concluding Remarks

In conclusion, the interdisciplinary dialogue between comprehensible input and perfuming offers a complementary interpretive perspective on several core issues in Krashen's Monitor Theory while highlighting the cognitive insights of the *Yogācāra* cognitive model. From the perspective developed in this study, language learning may be understood not merely as the accumulation of linguistic knowledge but as the gradual transformation of cognitive dispositions through repeated perfuming. This interdisciplinary framework may contribute to future research and pedagogical innovation in second language acquisition.

REFERENCES

- Abney, Steven P., and Mark Johnson. 1995. Memory requirements and local ambiguities of parsing strategies. *Journal of Psycholinguistic Research* 20:233–250.
- Dulay, Heidi C., and Marina K. Burt. 1974. Natural sequences in child second language acquisition. *Language Learning* 24(1):37–53.
- Ellis, Rod. 1994. *The Study of Second Language Acquisition*. Oxford: Oxford University Press.
- Ellis, Rod. 1997. *Second Language Acquisition*. Oxford: Oxford University Press.
- Gregg, Kevin R. 1984. Krashen's Monitor and Occam's Razor. *Applied Linguistics* 5(2):79–100.
- Gu, Qiyi. 2005. Implicit Knowledge, Explicit Knowledge and Disputes about Their Interface. *Foreign Language Education* 26(6):45–50.
- Horner, David. 1987. Misapplications of the Input Hypothesis. *ELT Journal* 41(3):177–183.
- Jiang, Jinlin. 2010. Disputable Issues on Explicit Learning and Implicit Learning. *Journal of Beijing Second Foreign Language University* 178(2):29–36.
- Jingjie. 2020. *Essentials of Yogācāra Studies*. Taipei: Spiritual Wisdom Publishing.
- Kochumuttom, Thomas A. 1982. *A Buddhist Doctrine of Experience: A New Translation and Interpretation of the Works of Vasubandhu the Yogācārin*. Delhi: Motilal Banarsidass.
- Krashen, Stephen D. 1981. *Second Language Acquisition and Second Language Learning*. Oxford: Pergamon Press.
- Krashen, Stephen D. 1982. *Principles and Practice in Second Language Acquisition*. Oxford: Pergamon Press.
- Krashen, Stephen D. 1985. *The Input Hypothesis: Issues and Implications*. Torrance, CA: Laredo Publishing Company.
- Krashen, Stephen D. 2004. *The Power of Reading: Insights from the Research*. 2nd edn. Westport, CT: Libraries Unlimited.
- Li, Gengchun. 2021. Frontiers and New Trends of Research on Implicit Learning in Second Language Acquisition. *Modern Foreign Languages (Bimonthly)* 46(4):576–584.
- Lin, Guoliang. 2007. *Cheng weishilun zhijie*. Shanghai: Fudan University Press.
- Long, Michael H. 1996. The role of the linguistic environment in second language acquisition. In *Handbook of Second Language Acquisition*, eds. William C. Ritchie and Tej K. Bhatia, pp.413–468. San Diego: Academic Press.
- Lusthaus, Dan. 2002. *Buddhist Phenomenology: A Philosophical Investigation of Yogācāra Buddhism and the Ch'eng Wei-shih Lun*. London: RoutledgeCurzon.
- McLaughlin, Barry. 1987. *Theories of Second-Language Learning*. London: Edward Arnold.
- Selinker, Larry. 1972. Interlanguage. *International Review of Applied Linguistics in Language Teaching* 10(1–4):209–232.
- Swain, Merrill. 1985. Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In *Input in Second Language*

- Acquisition*, eds. Susan M. Gass and Carolyn G. Madden, pp.235–253. Rowley, MA: Newbury House.
- Vasubandhu. *Trimśikā-vijñaptimātratā-kārikā* (Thirty Verses on Consciousness-Only).
- Waldron, William S. 2003. *The Buddhist Unconscious: The Ālaya-vijñāna in the Context of Indian Buddhist Thought*. London: RoutledgeCurzon.
- Wen, Qiufang. 2010. *Major Issues in Second Language Acquisition*. Beijing: Foreign Language Teaching and Research Press.
- Xuanzang. Baifa mingmen lun (Treatise on the One Hundred Dharmas). In *Taishō Shinshū Daizōkyō*, vol. 31, no. 1614. Edited by Junjirō Takakusu and Kaigyoku Watanabe. Tokyo: Taishō Issaikyō Kankōkai.
- Xuanzang (trans.). *Chéng Wéishì Lùn (Vijñaptimātratāsiddhi-śāstra)*. In *Taishō Shinshū Daizōkyō*, vol. 31, no. 1585. Edited by Junjirō Takakusu and Kaigyoku Watanabe. Tokyo: Taishō Issaikyō Kankōkai.
- Yang, Weizhong. 2014. *Consciousness-Only: A Buddhist Perspective*. Beijing: Religious Culture Press.
- Yijing. *Dàitáng xīyù qiúfǎ gāosēng chuàn* (Biographies of Eminent Monks of the Great Tang Seeking Scriptures in the Western Regions). In *Taishō Shinshū Daizōkyō*, vol. 51, no. 2066. Edited by Junjirō Takakusu and Kaigyoku Watanabe. Tokyo: Taishō Issaikyō Kankōkai.
- Yijing. *Nánhǎi jìguī nèifǎ chuán* (A Record of the Buddhist Religion as Practised in India and the Malay Archipelago). In *Taishō Shinshū Daizōkyō*, vol. 54, no. 2125. Edited by Junjirō Takakusu and Kaigyoku Watanabe. Tokyo: Taishō Issaikyō Kankōkai.
- Zeng, Yuping. 2009. On Interface between Explicit Knowledge and Tacit Knowledge in Second Language Acquisition. *Journal of Zhaoqing University* 38(4):38-42.

Wei-Wei Zhao
 Department of Chinese International Education
 Quanzhou Normal University
 Quanzhou, Fujian 362000, PRC
 cww6662@gmail.com

輸入與熏習：
Krashen 二語習得理論與佛教唯識學認知模型的跨學科對話

趙威維
泉州師範學院

本研究採用概念比較與文本分析的方法，探討第二語言習得理論中 Stephen Krashen(克拉申)監控理論與唯識學認知模型之間的概念對應與功能對應。研究聚焦於監控理論的五個假說，探討唯識學八識說與熏習種子理論是否能為監控理論中「輸入如何內化為語言能力」提供補充性的詮釋框架。本文提出，習得歷程可透過阿賴耶識的熏習與種子現行機制加以理解，而學習與監控功能則可與意識的分別作用進行比較分析。此外，本文以義淨與玄奘等佛教求法僧的語言學習經驗作為歷史例證，說明「輸入作為熏習」的可能意涵。研究認為，唯識學認知模型可作為理解語言內化與第二語言習得認知機制的補充性詮釋框架。

關鍵字：第二語言習得、監控理論、唯識學、八識、熏習、可理解輸入

TAILORING SPEAKING ACTIVITIES: THE IMPACT OF DIFFERENTIATED INSTRUCTION ON FIRST-YEAR UNIVERSITY STUDENTS

Adriani Yulia Purwaningrum
Universitas Negeri Yogyakarta, Indonesia

ABSTRACT

This study explores how Differentiated Instruction (DI) can transform speaking skills among first-year university students in English as a Foreign Language (EFL) setting. DI is implemented through carefully tailored speaking tasks designed to accommodate variations in students' proficiency levels and learning preferences. A total of 40 first-year university students participated in this classroom action research (CAR). Data were collected through a mixed-method approach. Quantitative evidence came from diagnostic assessment and speaking performances in two cycles, while qualitative data were obtained from classroom observations and students' reflection. Speaking performance was evaluated using a rubric measuring fluency, accuracy, and confidence. Findings suggest that DI fosters notable improvements in oral performance, with students demonstrating greater linguistic accuracy, smoother speech delivery, and increased willingness to participate in classroom discussions. Moreover, DI appears to boost student motivation and reduce speaking anxiety, creating a more inclusive and supportive learning environment. These results highlight DI as a promising approach for improving oral communication in diverse higher education classrooms, particularly in multilingual and multicultural settings.

Keywords: Differentiated Instruction, English as a Foreign Language (EFL), Foreign Language Anxiety, Speaking Skills

1. INTRODUCTION

Research has long indicated that English as a Foreign Language (EFL) learners often developed learning anxiety especially during speaking activities (Masrul and Erliana 2024; Hajiyeveva 2024; Saito et al. 2025; Tran 2022). This form of anxiety has been widely recognized as a significant barrier that hinders students' oral performance and participation. In response to this, various pedagogical strategies have been developed over the past few decades to reduce EFL learners' speaking anxiety (Hajiyeveva 2024; Kayhan 2025; Nguyen 2025; Özdemir and Seçkin 2025; Zarei and Rezadoust 2020; Zarin 2025). Among these, a growing body of literature has paid attention to the potential of Differentiated Instruction (DI) in addressing these challenges related to learner diversity, disproportionate participation, and speaking difficulties in EFL classrooms (Kurniawan et al. 2025; Oktavia 2025; Sapkota 2025; Zarei and Rezadoust 2020). The principle of DI, namely tailoring learning activities based on students' needs and interests, is a key to creating enjoyable learning experiences (Meşe and Mede 2023; Sapkota 2025; Tomlinson and Jarvis 2023). It is because DI acknowledges learner diversity and provides opportunities for students to engage with content in ways that match their readiness, interests, and learning profiles. By reducing the mismatch between instructional demands and learner characteristics, DI promotes optimal challenge, meaningful participation, and a sense of competence. These conditions foster positive learning experiences, which are closely associated with higher engagement, motivation, and achievement (Meşe and Mede 2023; Tomlinson and Jarvis 2023).

Provided that speaking is one of the most challenging productive skills for EFL learners to acquire, implementing student-centered approach such as DI becomes essential in supporting students' learning potential (Attia and Algazo 2025; Mujiono and Fatimah 2025; Nguyen 2025; Özdemir and Seçkin 2025). DI enables teachers to adapt instruction, tasks, and support according to learners' diverse needs and proficiency levels. One common instructional practice associated with DI is the provision of scaffolded feedback, which offers learners appropriate guidance based on their current proficiency level. Previous studies have reported that scaffolded

feedback during speaking activities help students navigate their own learning, resulting in significant improvements in speaking ability (Zarei and Rezadoust 2020). Furthermore, such tailored instructional practices have been shown to reduce students' anxiety by providing appropriate levels of support and challenge. However, despite the positive results, existing studies often do not explicitly address the diversity of learners within the classroom (Toyama and Yamazaki 2021; Zarin 2025). While DI is regarded as an effective approach, further research is needed to explore its efficacy in diverse classroom environments, where inclusivity remains a central objective.

In the Indonesian context, classrooms are inherently multicultural reflecting the country's vast diversity of around 694 languages spoken by more than 1,200 ethnic groups. This diversity contributes to a mixed-ability classroom settings, particularly in higher education, where students come from different provinces, linguistic communities, educational experiences, and socio-cultural backgrounds converge in the same classroom (Nursanti and Andriyanti 2021). Consequently, university EFL classrooms often exhibit substantial variation in learners' learning preferences, prior knowledge, language proficiency, and confidence in using English. As a result, such diversity presents additional challenges for educators in designing instructional practices that ensure equitable learning opportunities for all students (Alojado 2024; Liyanage and Tao, 2020; Raihani 2018).

Despite the increasing interest in DI, most existing studies in Indonesia have primarily focused on its implementation at the secondary education level and provided limited information regarding learners' linguistic and sociocultural diversity (Purwaningrum et al. 2025; Religioni et al. 2024; Sapkota 2025; Tomlinson et al. 2003). Although these studies acknowledge differences in students' proficiency levels, they provide limited insights into how DI operates in university classrooms with broader linguistic, academic, and cultural diversity. Furthermore, there is still a lack of studies that specifically investigate the use of DI in tailored speaking activities supported by scaffolding, especially among first-year university students who are transitioning from varied educational backgrounds and often experience differing levels of speaking anxiety

(Oktavia 2025). This gap highlights the need to investigate how DI can support speaking development in diverse higher education EFL contexts.

Therefore, this study aims to investigate the implementation of Differentiated Instruction (DI) in enhancing first-year university students' speaking skills within a diverse EFL classroom. Specifically, this study seeks to examine how tailored speaking activities supported by scaffolding can improve students' fluency, accuracy, and confidence, while also addressing their speaking anxiety.

2. LITERATURE REVIEW

2.1 Differentiated Instruction in EFL Speaking Classrooms

Differentiated Instruction (DI) is a pedagogical approach that emphasizes adapting teaching methods, materials, and learning activities to accommodate learners' diverse needs, proficiency levels, and learning preferences (Gheyssens et al. 2020; Tomlinson 2017; Tomlinson and Jarvis 2023). In EFL contexts, where classrooms are often characterized by mixed abilities and varied linguistic backgrounds, DI plays a crucial role in ensuring inclusive and equitable learning opportunities. Research indicates that DI involves modifying content, process, and output to engage students more effectively and support their individual learning trajectories (Mujiono and Fatimah 2025; Purwaningrum et al. 2025; Subban et al. 2025). DI allows students to participate in speaking activities at an appropriate level of challenge by adapting instructional materials, learning activities, and assessment to match their readiness levels, interests, and learning profiles. Such adaptations have been correlated to the increased participation, confidence, and language development.

In speaking instruction, DI is particularly beneficial as it allows teachers to design tailored speaking tasks that align with learners' readiness levels (Ningsih and Pusparini 2024). Empirical studies have shown that differentiated speaking activities, such as tiered tasks, flexible grouping, and varied interaction formats, can significantly enhance students' oral performance. For example, Oktavia (2025) found that DI improves students' speaking ability by enabling them to participate in tasks suited to their proficiency levels, resulting in better fluency and engagement. Similarly, Saefullah et al. (2025) reported that integrating DI

with communicative activities fosters active participation and supports speaking development.

In addition to performance improvement, DI has also been associated with positive affective outcomes. According to Saputro and Sahiruddin (2025), students who participated in learning activities aligned with their readiness levels reported higher motivation and reduced anxiety. These findings suggest that differentiated learning experiences can contribute to both cognitive and affective development.

However, despite these promising findings, several gaps remain. First, many studies on DI in EFL contexts have primarily focused on general language proficiency rather than specific speaking components such as fluency, accuracy, and confidence (Ningsih and Pusparini 2024). Second, most research has been conducted in secondary school settings, with limited attention to higher education contexts (Ningsih and Pusparini 2024; Saefullah et al. 2025). Third, although DI emphasizes inclusivity, most existing studies primarily addressed proficiency-related diversity and provided limited information regarding linguistic, cultural, or regional diversity. Such studies provide limited insight into university classrooms that accommodate students from different regions, ethnic groups, first-language backgrounds, and prior educational experiences. Consequently, evidence regarding how DI can address multiple dimensions of learner diversity in higher education EFL classrooms remains limited. These gaps highlight the need for further investigation into how DI can be effectively applied in university-level speaking classrooms.

2.2 Speaking Anxiety and the Role of Scaffolding in Differentiated Learning

Speaking anxiety is widely recognized as one of the most significant challenges faced by EFL learners. It is often associated with fear of negative evaluation, lack of confidence, and limited linguistic competence, all of which can hinder students' willingness to participate in speaking activities. Research consistently shows that high levels of anxiety negatively affect speaking performance, leading to reduced fluency, avoidance of interaction, and lower achievement (Cahya and Churnia 2026).

To address this issue, scaffolding has been widely adopted as an instructional strategy that provides learners with temporary support to accomplish tasks beyond their current ability (Zarei and Rezadoust, 2020). In speaking instruction, scaffolding may include modelling, guided practice, structured prompts, and feedback. Studies have demonstrated that scaffolded instruction significantly improves students' speaking performance by enhancing both fluency and accuracy while gradually building learners' confidence (Kayhan 2025; Özdemir and Seçkin 2025). Furthermore, scaffolded feedback has been found to reduce speaking anxiety by helping learners feel more supported and less fearful of making mistakes (Zarei and Rezadoust 2020).

Recent research also suggests that scaffolding and DI are highly complementary approaches (Purwaningrum et al. 2025). While DI focuses on tailoring instruction to learners' needs, scaffolding provides the necessary support to help learners successfully complete differentiated tasks. For instance, Ghaith and Awada (2022) highlight that combining DI with scaffolding strategies can mitigate learners' anxiety and promote deeper engagement. Similarly, Huang (2024) found that scaffolded learning environments contribute to reduced anxiety and improved speaking performance in EFL contexts.

Nevertheless, important gaps remain in the literature. Although previous studies have examined speaking anxiety and scaffolding independently, fewer studies have explored their interaction within a Differentiated Instruction framework, particularly in speaking classrooms. In addition, most research tends to focus on either affective factors (e.g., anxiety) or performance outcomes (e.g., fluency), rather than examining both dimensions simultaneously. Existing studies have also been conducted predominantly in primary and secondary educational contexts (Religioni et al. 2024; Sapkota 2025), with comparatively limited research in higher education settings. Furthermore, limited attention has been given to how students perceive and respond to differentiated and scaffolded speaking activities, especially in multicultural university settings. Therefore, further research is needed to investigate the relationship between DI-supported scaffolding, students' speaking performance, speaking anxiety, and learning experiences in diverse university EFL classrooms.

2.3 Research Questions

1. How does differentiated instruction affect freshmen's speaking performance in English?
2. What changes occur in students' speaking confidence and anxiety after participating in DI-based speaking activities?
3. How do students perceive differentiated speaking activities in terms of engagement and learning value?

3. METHODOLOGY

3.1 Research Design

This study employed a Classroom Action Research (CAR) design, focusing on enhancing first-year students' speaking proficiency through the implementation of differentiated instruction. The cyclical model of planning, acting, observing and reflecting by Burns (2024) was adopted as it enables the researcher to systematically evaluate instructional interventions while making ongoing adjustments based on the evidence in classroom practice. In this case, this study collected both quantitative and qualitative data to obtain a comprehensive understanding of the intervention. Quantitative data were collected to examine changes in students' speaking performance and anxiety levels across the intervention. Meanwhile, qualitative data provided in-depth understanding of students' perceptions, learning experiences, and responses to differentiated speaking activities. These data sources were integrated to provide triangulation, which makes the findings more reliable and valid since they captured not only measurable outcomes, but also contextual classroom processes.

The CAR was conducted in two cycles, each consisting of six meetings, with one meeting dedicated for reflection between the cycles. The use of two cycles allowed the researcher to identify challenges and areas for improvement during the first cycle and subsequently refine instructional strategies in the second cycle. This process is beneficial to verify whether the changes observed during the learning were sustained after revisions were implemented, which enhances the reliability and trustworthiness of

the findings. Consistent with CAR principles, the second cycle also provided an opportunity to validate reflections from the first cycle and assess the effectiveness of the modified instructional actions.

The intervention was implemented in a pre-intermediate oral communication skills course aimed at developing students' ability to communicate in English with greater fluency, accuracy, interaction, and coherence at the A2 CEFR level. The DI was implemented through flexible grouping, varied speaking tasks, and multiple tasks options that accommodated students' proficiency levels and learning preferences. Speaking activities included self-introduction, descriptions of people, places, activities, and objects, invitation-based conversation, discussions, and role-play. Scaffolding was provided through vocabulary support, guiding questions, structured practice, peer collaboration, and formative feedback. This intervention was conducted over two cycles comprising twelve instructional meetings, with each meeting lasting approximately 100 minutes.

3.2 Participants

The participants were 40 first-year university students undertaking pre-intermediate oral communication skills course at a public university in Yogyakarta, Indonesia. The participants were selected because previous research in Indonesian higher education contexts reported that first-year EFL students often struggle in speaking and demonstrate low engagement in speaking class (Kusumawati 2020). The participants were enrolled in 2 (two) different classes, and both represented mixed-ability groups based on variations in their speaking proficiency and learning needs, which are commonly observed in Indonesian classrooms. Individual consent was obtained prior to data collection. However, since this study focused primarily on learner diversity in terms of proficiency and instructional needs, detailed demographic data related to their linguistic or cultural backgrounds were not collected.

3.3 Data Collection Procedure

To ensure triangulation, data were collected using both quantitative and qualitative instruments. Both instruments were designed to measure the speaking performance and speaking anxiety level of the students.

Students' speaking performance was assessed through three speaking assessments administered at different stages of the intervention. The diagnostic test required students to introduce themselves and provide simple description using one image of their choice. The progress test assessed students' ability to perform more comprehensive self-introductions, describe people, places, activities, and objects, as well as engage in invitation-related speaking tasks. The post-test was conducted at the end of Cycle 2, which consisted of a role-play activity that required students to integrate and apply the speaking functions and language features covered throughout the course. The assessment rubric for all tasks was aligned with The Common European Framework of Reference for Languages (CEFR) oral assessment criteria including range, accuracy, fluency, interaction, and coherence set for A2 level (pre-intermediate) learners. Students' speaking performances were independently assessed by the researcher, who also served as the course lecturer, and a collaborating lecturer using the same assessment rubric. Prior to scoring, both raters reviewed the rubric descriptors to ensure consistency in scoring. Following recent recommendation for classroom-based assessment, reliability test was measured using Cronbach's Alpha.

Meanwhile, the students' speaking anxiety level was measured using Foreign Language Classroom Anxiety Scale (FLCAS) adapting Elaine K. Horwitz (1986). While the original FLCAS consists of 33 items, this study simplified the instruments into a 10-item questionnaire to reduce participant burden and improve its practicality within the classroom action research setting. The researcher ensured that selected items remained representing three core dimensions of foreign language anxiety, namely communication apprehension, test anxiety, and fear of negative evaluation. Moreover, 2 (two) items were specifically designed to be reverse coded to minimize bias.

Additionally, qualitative data were collected from classroom observations and students' reflective journals. A checklist was prepared to guide the observation, following DI framework and anxiety indicators. The observations focused on students' participation, collaboration, classroom engagement, and observable signs of anxiety during speaking activities. The observations were conducted by the collaborating lecturer, and field notes were recorded after each session to document important

classroom episodes and students' responses to the instructional activities. In addition, students were asked to complete a reflective journal as an exit ticket at the end of each cycle. The reflections were specifically designed to offer opportunities for the students to describe their learning experiences and perceptions of the differentiated speaking activities. Students were also instructed to reflect on to any challenges they experienced during the learning process. In this way, the qualitative data provided contextual insights that complemented the quantitative findings.

Table 1. Sample of observation checklist and field note

	Checklist Items	Yes	No	Side notes
Differentiation Content	Students at different proficiency levels can follow the material.			
Differentiation Process	Students collaborate well in flexible groups (pair, small group, mixed-level).			
Differentiation Product	Students complete speaking tasks at different complexity levels.			
Student Emotional Responses & Anxiety Indicators	Students volunteer to speak without being called.			

...

3.4 Data analysis

Quantitative data were analyzed statistically using SPSS (Statistical Package for the Social Sciences) software. To analyze changes across three time points (diagnostic, cycle 1, and cycle 2), normality test was conducted. It was followed by a repeated-measures ANOVA to determine significant differences in speaking performance and anxiety levels across the three stages of the intervention. Post hoc pairwise comparisons with Bonferroni adjustments were applied to identify where the improvements occurred. Effect sizes (partial eta squared) were also calculated to measure the magnitude of the intervention's impact.

To complement these analyses, descriptive statistics, including means, standard deviations, and percentage changes, were computed to show overall trends and the direction of improvement. Visualizations such as line graphs were used to illustrate the progression of students' speaking scores and anxiety levels from the diagnostic test to Cycle 2. These quantitative procedures provided statistical evidence of the effectiveness of differentiated instruction in enhancing speaking performance and reducing language anxiety.

Meanwhile, thematic analysis approach was employed to analyze the qualitative data. This type of analysis technique was used owing to its flexibility and practicality to build up a constructionist and critical analysis that helps identifying the concepts and ideas of the content (Braun and Clarke 2023). Following the analysis steps by Braun and Clarke, the data were first read to gain familiarity and identify meaningful units of information. Then, these units were coded based on patterns related to students' engagement, use of differentiated tasks, collaboration, confidence indication, and behavioral signs of reduced anxiety during speaking activities. Afterwards, the similar codes were grouped into broader themes that described changes in students' behavior and emotional readiness to speak. The findings were then compared with the quantitative results to provide a richer understanding of how and why differentiated instruction contributed to improvements in speaking performance and reductions in speaking anxiety.

Table 2. Display of thematic analysis

Theme	Code Examples	Description / Interpretation	Cycle Evidence
Reduced Speaking Anxiety	long pause, avoids speaking, softer voice, relaxed posture, fewer pauses, attempts without prompting	Students gradually became less tense and more willing to speak.	Pre → C1 → C2
Increased Engagement	initiates task, stays on task, asks questions, participates actively	Students showed more active participation and task involvement over cycles.	C1 → C2
...			

4. FINDINGS

4.1 Quantitative Data

The findings of this study present the effects of differentiated instruction (DI) on students' speaking performance across three stages of classroom action research: the diagnostic pre-test, Cycle 1, and Cycle 2. In addition, it also explores the change of students' speaking anxiety before and after the intervention. Quantitative analyses included descriptive statistics, tests of normality, a repeated-measures ANOVA to identify changes of speaking performance across time and paired-sample t-Test to examine changes in speaking anxiety.

Table 3. Descriptive Statistics

		N	Min.	Max.	Mean	SD
Speaking Performance	Diagnostic	40	50	70	59.5	5.86
	Cycle 1	40	60	85	70.75	6.36
	Cycle 2	40	60	90	73.3	8.21
Speaking Anxiety	Pre-test	40	-	-	2.78	0.43
	Post-test	40	-	-	2.61	0.40

Table 3 presents the descriptive analysis of the data on speaking performance across the three measurements, showing a clear upward trend of the score. The mean for diagnostic test was 59.5 (SD=5.86), reflecting students' initial speaking ability. After implementing DI in Cycle 1, the mean increased to 70.75 (SD = 6.36), and a further increase was observed in Cycle 2 with a mean of 73.3 (SD = 8.21). The improvement from Cycle 1 to Cycle 2 indicates that students continued to benefit from differentiated learning tasks and scaffolding strategies throughout the intervention.

In addition, it also presents the output for FLCAS before and after the differentiated instruction intervention. As can be seen in Table 2, students reported higher anxiety levels in the pre-test (mean=2.78; SD= 0.43) compared to the post-test (mean= 2.61; SD= 0.40). It suggests that students possess a moderately high level of speaking anxiety at the beginning of the study but decreases significantly after the intervention. However, the relatively similar standard deviations across the two measurements indicate that students varied in their anxiety levels to a comparable degree at both time points.

Table 4. Normality Test

	Statistic	df	Sig.
--	-----------	----	------

Diagnostic	.914	40	.005
Cycle 1	.895	40	.001
Cycle 2	.951	40	.084

*Shapiro Wilk

Normality was assessed using the Shapiro–Wilk test because the sample size was below 50 ($n=40$). As shown in Table 3, the diagnostic data ($W = .914$, $p = .005$) and Cycle 1 data ($W = .895$, $p = .001$) did not meet the assumption of normality, whereas the Cycle 2 data were normally distributed ($W = .951$, $p = .084$). Although violations of normality were observed in two measurement points, repeated-measures ANOVA was retained because the procedure is generally robust to moderate departures from normality, particularly in repeated-measures designed with balanced data and moderate sample sizes. Therefore, the analysis proceeded with repeated-measures ANOVA while additional assumption checks were conducted.

Mauchly's Test of Sphericity was conducted to evaluate the assumption of equal variances of the differences between levels of the within-subjects factor (Time). The test was significant, $\chi^2(2) = 14.75$, $p = .001$, indicating that the sphericity assumption was violated. Therefore, the Greenhouse–Geisser correction ($\epsilon = .757$) was applied to adjust the degrees of freedom.

In this case, a repeated-measures ANOVA was conducted to examine changes in speaking performance across three time points. All multivariate tests indicated a significant main effect of Time, Pillai's Trace = .806, $F(2, 38) = 78.92$, $p < .001$, partial $\eta^2 = .806$.

This analysis suggests a very large effect size, meaning that 80.6% of the variance in scores was explained by changes over time. Thus, it confirms that students' speaking performance changed significantly from the diagnostic test through Cycle 1 to Cycle 2.

In addition to analyzing changes in speaking performance using repeated-measures ANOVA, a paired-samples t-test was also conducted

to examine changes in speaking anxiety comparing FLCAS pre- and post-test scores.

Table 5. Paired Samples Test

	Mean	t	df	Sig. (2-tailed)
Pair 1	.17056	3.578	39	.001

Table 5 shows that $p = .001$, which is $< .05$, indicating that the difference between pre- and post-anxiety levels is statistically significant. This finding infers that DI contributed meaningfully to reducing students' foreign language anxiety.

Overall, the quantitative results demonstrate a significant and substantial improvement in students' performance across the three measurement times, as well as a significant decrease in students' speaking anxiety. The moderate effect on FLCAS combined with the very large effect on speaking performance suggests that while DI strongly enhanced skill development, changes in affective variables were meaningful but more gradual.

4.2 Qualitative Data

The thematic analysis generated several recurring patterns that reflected students' behavioral, emotional, and cognitive responses to the differentiated speaking activities. Codes that showed reluctance to speak, long pauses, or avoidance during early sessions evolved into clustered patterns describing reduced tension and increased willingness to participate, forming the theme *Reduced Speaking Anxiety*. Similarly, codes related to frequent group interaction, active task completion, and voluntary participation were organized into the theme *Increased Engagement*. Other patterns, such as students helping peers, negotiating meaning, and interacting within flexible groupings, led to the theme *Improved Collaboration*. Codes capturing students' independent planning,

use of strategies, and ability to self-correct contributed to the theme *Greater Learner Autonomy*. Finally, consistent improvements in voice clarity, idea organization, and readiness to speak resulted in the theme *Growth in Speaking Confidence*. These themes collectively illustrate how differentiated instruction gradually transformed the speaking environment and student behavior from the diagnostic stage through Cycle 1 and 2.

Table 6. Display of thematic analysis

Theme	Sub-Themes	Codes (Condensed Meaning Units)	Evidence From Data (Observations and students' checklist)	Interpretation
1. Reduced Speaking Anxiety	- Scaffolding reduces fear - Structured tasks build security	- Hesitation decreases when sentence frames provided - Students show fewer signs of nervousness - Participation increases when supported	<i>"Students seemed calmer when the teacher provided guided prompts."</i> <i>"Fewer long pauses observed during structured pair work."</i> (observation) <i>"I feel less insecure and I can understand what is being learned."</i> (student's reflection)	Differentiated scaffolding lowers psychological pressure; students gain confidence when linguistic and procedural supports are available.

2. Increased Engagement	- Interest-based tasks increase motivation	- Students participate more in tasks they choose	<i>"Most students volunteered to present when allowed to choose topics."</i>	Offering choice and varied difficulty increases autonomy and intrinsic motivation, leading to more consistent engagement.
	- Flexible grouping encourages participation	- More active discussion in small groups - Increased willingness to collaborate	<i>"Groups showed lively interaction during choice-based speaking tasks."</i> (observation)	
			<i>"I like the different approach because I can have a different experience and it's not monotonous."</i> (student's reflection)	
3. Growth in Speaking Confidence Across Cycles	- Gradual improvement in fluency	- Longer utterances in Cycle 2	<i>"Several students initiated conversation without waiting for the teacher."</i>	Confidence strengthened over time; differentiated activities enabled students to progress
	- Greater willingness to initiate speech	- More self-initiated responses	<i>"Students produced longer,</i>	

		- Reduced reliance on scripts	<i>more natural sentences in Cycle 2.</i> " (observation)	ly expand their speaking competence
			<i>"Now I feel more confident speaking in front of my classmates."</i> (student's reflection)	
4. Improved Peer Collaboration and Interaction	- Supportive peer dynamics - Positive group atmosphere	- Students help peers with vocabulary - Increased peer encouragement - Smoother turn-taking	<i>"Students corrected one another gently during group tasks."</i> <i>"Group members encouraged shy students to speak."</i> (observation)	Differentiated grouping fostered a supportive environment that helped reduce anxiety and improve fluency.
			<i>"I feel more comfortable and confident with other friends when speaking."</i> (student's reflection)	

5. Greater Autonomy in Completing Speaking Tasks	- Reduced teacher dependency - Emergence of independent strategies	- Students negotiate meaning on their own - Use of self-monitoring strategies - Students manage task pacing	<i>“Students used gestures and paraphrasing to clarify meaning independently.”</i> <i>“Teacher intervention reduced significantly in Cycle 2.”</i> (observation) <i>“I can prepare my thoughts, work on my pronunciation, and express myself more freely through video projects.”</i> (student’s reflection)	Students became more autonomous learners; differentiation encouraged them to take ownership of their speaking process.
---	---	---	---	---

5. DISCUSSION

This section interprets the findings of this study to answer the research questions. Students’ speaking performance is discussed from how the score improved within the three periods of the study. Meanwhile, students’ behavioral changes cover the transformations of the students during the study. It was observed from the students’ engagement, speaking anxiety level, speaking confidence, and autonomy during the learning process. Finally, students’ perceptions were reported based on their reflections before, during, and after the lesson.

5.1 Students' speaking performance

The quantitative findings demonstrated a clear improvement in students' speaking performance following the implementation of differentiated instruction. Results from the repeated-measures ANOVA showed a progressive increase across the three testing points, indicating that students' speaking skills developed gradually rather than abruptly. This pattern suggests that the intervention was not only impactful but also sustained over time. The steady improvement in mean scores indicates positive improvement in key speaking components, including fluency, accuracy, and communicative confidence.

Moreover, the qualitative findings suggest that the differentiated and scaffolded speaking activities contributed to the observed improvement in the students' speaking performance. The intervention, including provision of tasks and autonomy to present the learning products that were tailored to students' needs and proficiency level, enabled the students to participate in speaking activities at an appropriate level of difficulties. The scaffolding embedded within the speaking activities was prone to encourage students in organizing ideas, accessing relevant vocabulary, and maintaining communication during speaking tasks. Consequently, students were able to pay more attention to meaning making and interaction, instead of struggling with linguistic form and content generation all together.

In addition, the qualitative findings offer further insights into the means underlying this improvement. Many students reported that the differentiated activities helped them to "organize thoughts", "avoid freezing", and "remember vocabulary better" during speaking activities. Others reported that learning speaking with activities tailored to their learning preferences and proficiency levels increased their confidence and willingness to participate. These reflections signified that the intervention not only provided additional opportunities for speaking practice but also encouraged a more supportive learning environment in which students felt more comfortable taking risks and expressing their ideas.

This interpretation is aligned with previous research indicating that targeted strategy instruction can support learners in becoming more

autonomous speakers and more intentional in planning, monitoring, and evaluating their speech (Ningsih and Pusparini 2024). However, the present study broadens this understanding by providing evidence on how differentiated activities, combined with scaffolding, can reinforce speaking development in a mixed-ability university EFL classroom. The combination of quantitative improvements and students' experiences reported in this study reveals several factors affecting students' speaking development. In addition to the increased practice opportunities, it was also influenced by the students' growing confidence, strategic awareness, and engagement during the intervention. Collectively, both datasets point to a positive relationship between the intervention and students' observable speaking progression.

5.2 Students' behavioral changes

Behavioral changes became particularly visible during the implementation. Students gradually shifted from being passive recipients of instruction to more active participants in communicative tasks. It was supported by the quantitative findings, as FLCAS scores showed a noteworthy reduction in students' speaking anxiety following the intervention as confirmed by the paired-samples t-test ($t(39) = 3.578$, $p = .001$). This reduction was consistent with students' reflections, which signaled increased confidence and contentment during speaking activities. For instance, one student testified, "I feel less insecure, and I can understand what is being learned," while another expressed that differentiated activities "make me more confident to express myself." Correspondingly, students valued opportunities to learn according to their learning preferences and proficiency levels, with one student stating that "when we (were) divided into groups, I felt confident." These evidence from the findings reveal that students perceived the differentiated activities and learning environment facilitated students to feel more comfortable engaging in speaking activities by increasing tolerance for ambiguity and reducing fear of making mistakes, which were common factors associated with foreign language anxiety (Horwitz 1986).

Furthermore, the qualitative responses revealed additional behavioral changes that extended beyond measurable performance. Students

frequently reported becoming more willing to speak, take risks, and initiate contributions during group work. Many also expressed feeling more confident because they were given appropriate levels of support and flexibility in completing speaking tasks. These perceptions imply that differentiated instruction established a learning environment where students felt psychologically safe to experiment with language and take communicative risks.

Another prominent finding was the development of self-regulatory behaviors. Students reported to plan their responses more carefully, prepare vocabulary before speaking tasks, and evaluate their performance afterward. These behaviors are characteristics of metacognitive strategy use, which involves planning, monitoring, and evaluating student's learning processes. Suharto et al. (2025) reported that these metacognitive processes positively influence both performance and confidence by helping students become more aware of their strengths and areas of improvement. Thus, these behavioral shifts reflect not only reduced anxiety but also increased learner agency and ownership of the learning process.

These findings extend previous research by highlighting the role of affective and behavioral factors in supporting students' speaking development within differentiated learning environment. The increase of students' speaking performance came along with reduced anxiety, greater participation, and stronger self-regulatory behaviors. It indicated that students' willingness and readiness to engage in communication are among important mechanisms on how differentiated instruction facilitates oral language development.

5.3 Students' perceptions

Students' perceptions of the intervention were overwhelmingly positive. Based on the thematic analysis, they valued the personalized and flexible nature of the differentiated approach. Many students expressed that the strategies felt "manageable," "relevant," and "easy to apply." They appreciated that the intervention acknowledged their individual needs rather than treating them as a uniform group. This perception may

help explain the improvements in motivation and participation observed during the program.

Another strong theme was relevance to real communication. Students felt that the strategies equipped them with tools applicable beyond the classroom, such as planning responses in everyday English use or managing anxiety in spontaneous exchanges. Some students also acknowledged improvement in self-confidence, which they attributed to repeated practice and scaffolded speaking activities (Huang 2024; Zarei and Rezadoust 2020). These perceptions reinforce the quantitative improvements and highlight the subjective value students place on instruction that feels meaningful and personally applicable.

In general, the findings suggest that differentiated instruction approach positively influenced students' speaking development, behavioral engagement, and emotional responses toward speaking English. Improvements were not only measurable in performance and anxiety levels but also visible in how students approached learning and engaged with communicative tasks. The convergence of quantitative and qualitative evidence indicates that the intervention was effective and meaningful from both performance-based and learner-experience perspectives.

6. CONCLUSION

This study investigated the implementation of Differentiated Instruction (DI) supported by scaffolding in a first-year university EFL speaking classroom. The findings reveal that students demonstrated gradual improvement in speaking performance across the intervention cycles, along with the reduced speaking anxiety and positive perceptions of the learning experience. Collectively, both quantitative and qualitative findings imply that tailored and scaffolded speaking activities support both linguistic development and affective aspects of language learning.

Taken together, this study makes several contributions to the current literature in several ways. First, it broadens research on Differentiated Instruction by scrutinizing its application in a university-level EFL speaking context, an area which was underrepresented in previous studies.

Second, it presents empirical evidence related to the blended use of scaffolding and differentiation to support speaking development while addressing speaking anxiety. Third, it offers a more comprehensive understanding regarding the learners' experience and respond to differentiated speaking instruction by integrating quantitative measures of speaking performance and anxiety with qualitative insights from observations and reflections.

Furthermore, from a pedagogical perspective, this study indicates that EFL teachers may benefit from designing speaking activities that accommodate students' different proficiency levels, learning preferences, and support needs. This tailored instruction is beneficial to foster greater participation, confidence, and engagement in speaking classrooms. Overall, this study highlights the potential of differentiated and scaffolded instruction as a student-centered approach for enhancing speaking development in higher education EFL contexts.

7. LIMITATIONS AND FUTURE STUDIES

Several limitations should be considered when interpreting the findings of this study. First, the small sample size of only 40 first-year university students from a single public university in Indonesia did not allow generalizability of the findings to other educational contexts or learner populations. Second, the most important limitation lies in the fact that the qualitative data partially relied on students' self-reported reflections and responses to FLCAS. Although these instruments provided valuable insights into students' perceived understanding and emotional experiences, self-reported data are subject to social desirability bias and individual differences in self-awareness. Third, one source of weakness in this study which could have affected the measurements of students' speaking performance was the intervention that was only conducted over two cycles within a single semester. Even though the findings indicate positive improvements over time, a longer period of intervention may provide a more comprehensive understanding of the sustainability of the observed changes in speaking performance, anxiety, and learner behaviors.

Finally, the scope of this study was limited in terms of data sources, which only focused on speaking assessments, classroom observations, and

student reflections. It is unfortunate that the study did not include a systematic analysis of students' learning artefacts, such as speaking recordings, portfolios, or task products. The inclusion of such artefacts could provide deeper insights into the processes underlying students' speaking development and strategy use.

In this light, future studies should address these limitations to establish a greater impact on this matter. It is suggested that further research can involve larger and more diverse participant groups, extend the duration of the intervention, combine more data sources, and examine the implementation of differentiated and scaffolded instruction across different EFL contexts and proficiency levels.

REFERENCES

- Alojado, Ginnalyn. 2024. Teachers' best practices in meeting diverse learners' needs: An exploratory sequential design. *Southeast Asian Journal of Multidisciplinary Studies* 4.
- Attia, Shaden., and Muath Algazo. 2025. Foreign language anxiety in EFL classrooms: teachers' perceptions, challenges, and strategies for mitigation. *Frontiers in Education* 10: 1614353.
- Braun, Virginia., and Victoria Clarke. 2023. Toward good practice in thematic analysis: Avoiding common problems and being a knowing researcher. *International Journal of Transgender Health* 24: 1–6.
- Burns, Anne. 2024. *Theoretical framework: Second language teacher education and action research*. The Routledge Handbook of Language Teacher Action Research. New York: Routledge.
- Cahya, Nur., and Elfi Churnia. 2026. Communication anxiety and speaking performance in EFL learners: A systematic review of counseling and pedagogical strategies. *In The Indonesian Journal of Educational Counseling* 10.
- Ghaith, Ghazi M., and Ghada M. Awada. 2022. Scaffolding understanding of scholarly educational research through teacher/student conferencing and differentiated instruction. *Teaching and Learning Inquiry* 10.
- Gheysens, Esther., Catherine Coubergs., Julia Griful-Freixenet., Nadine Engels., and Katrien Struyven. 2020. Differentiated instruction: the diversity of teachers' philosophy and praxis to adapt teaching to students' interests, readiness and learning profiles. *International Journal of Inclusive Education*.
- Hajiyeva, Bulbul. 2024. Language anxiety in ESL learners: Causes, effects, and mitigation strategies. *EuroGlobal Journal of Linguistics and Language Education* 1: 119–133.
- Horwitz, Elaine K. (1986). Preliminary evidence for the reliability and validity of a foreign language anxiety scale. *TESOL Quarterly* 20: 559–562.
- Huang, Hsiang Ling. 2024. Exploring the impact of VR scaffolding on EFL teaching and learning: Anxiety reduction, perceptions, and influencing factors. *Multimodal Technologies and Interaction* 8: 85.
- Kayhan, Selvanur. 2025. The effect of intervention studies on foreign language speaking anxiety: A meta-analysis study. *Journal of Pedagogical Research* 9: 62–83.
- Kurniawan, Ildi., Azhar Aziz Lubis., and Eka Novita. 2025. The implementation of Differentiated Instruction to improve students' speaking skills in an academic speaking class. *Journal of English Language Studies* 10: 334–349.
- Kusumawati, Anggara Jatu. 2020. Redesigning face-to-face into online learning for speaking competence during Covid-19: ESP for higher education in Indonesia. *International Journal of Language Education* 4: 276–288.
- Liyanage, Indika., and Wenhao Tao. 2020. Preparation of teachers and multilingual education: Ethical, just, and student-focused practices. *In Multilingual Education Yearbook 2020: Teacher Education and Multilingual Context*. Springer.

- Masrul., Santi Erliana. 2024. The influence of language learning anxiety on speaking and reading performances in foreign language context. *Study of Applied Linguistics and English Education* 1.
- Meşe, Esra., and Enisa Mede. 2023. Using digital differentiation to improve EFL achievement and self-regulation of tertiary learners: The Turkish context. *Innovation in Language Learning and Teaching* 17: 340–353.
- Mujiono, Mujiono., and Siti Fatimah. 2025. Evaluating the impact of Differentiated Instruction on English language proficiency in EFL classrooms: A meta-analytic perspective. *Jurnal Pendidikan Progresif* 15:2328–2347.
- Nguyen, Thuc Anh. 2025. Addressing English-speaking anxiety in Vietnamese initial teacher education: Beyond coping strategies. *Asian-Pacific Journal of Second and Foreign Language Education* 10: 24.
- Ningsih, Ratna., and Ririn Pusparini. 2024. Differentiated learning in teaching speaking: Challenges and solutions. *ELTIN Journal: Journal of English Language Teaching in Indonesia* 2.
- Nursanti, Emi., and Erna Andriyanti. 2021. Language identities of multilingual college English learners in Indonesia. *Eurasian Journal of Applied Linguistics* 7:316–337.
- Oktavia, Nelly. 2025. The effectiveness of differentiated instruction based on learning styles in improving EFL speaking performance. *Journal of English as a Foreign Language Education (JEFLE)* 6: 79–88.
- Özdemir, Osman., and Hümset Seçkin. 2025. Exploring foreign language anxiety in higher education: Multifaceted insights into causes, impacts, and coping strategies. *Social Sciences & Humanities Open* 11:101364.
- Purwaningrum, Adriani Yulia., Margana, and Nur Hidayanto Pancoro Setyo Putro. 2025. Enquiry towards middle-school teachers about Differentiated English Instruction in a big classroom. *AsTEN Journal of Teacher Education* 9:9.
- Raihani, R. 2018. Education for Multicultural Citizens in Indonesia: Policies and Practices. *Compare* 48:992–1009.
- Religioni, Akifayadia., Calista Meilinda., Asri Alawiyah., and Luka Muhammad Farrel. 2024. Differentiated instructions in ESL and EFL classrooms: A systematic literature review. *STAIRS: English Language Education Journal* 5.
- Saefullah, Asep., Mashadi Said., and Venny E. Meidasari. 2025. The implementation of Differentiated Instruction and task-based instruction in teaching speaking skills within the Merdeka curriculum (Study at junior high schools of South Tangerang). *INFERENCE: Journal of English Language Teaching* 8: 188–201.
- Saito, Kazuya., Jean-Marc Dewaele., Yo In'nami., and Mariko Abe. (2025). Disentangling the causal role of motivation, enjoyment, and anxiety in second language speech learning: A final report. *Studies in Second Language Acquisition* 47:461–487.
- Sapkota, Hari Prasad. 2025. Differentiated Instruction in EFL classrooms: A systematic review. *Linguistics (JNLL)* 3.
- Saputro, Hana Sabrina Putri Khairunnisa., and Sahiruddin. 2025. Speaking anxiety and factors in Indonesian EFL tertiary contexts: Exploring proficiency differences. *International Journal of Language and Literary Studies* 7:126–138.

- Subban, Pearl., Muhamad Nanang Suprayogi., Melly Preston., Annisa Nanda Liyani., and Andrea Prita Purnama Ratri. 2025. "Differentiation is sometimes a hit and miss". Educator perceptions of differentiated instruction in the higher education sector. *The Asia-Pacific Education Researcher* 34:873–884.
- Suharto, Pipit Prihartanti., Ika Lestari Damayanti., and Nenden Sri Lengkanawati. 2025. Exploring metacognitive strategies to support young learners in developing their learner autonomy. *International Journal of Language Education* 9:331–355.
- Tomlinson, Carol Ann. 2017. *How to differentiate instruction in academically diverse classrooms* (3rd ed.). ASCD.
- Tomlinson, Carol Ann., Catherine Brighton., Holly Hertberg., Carolyn M. Callahan., Tonya R. Moon., Kay Brimijoin., Lynda A. Conover., and Timothy Reynolds. 2003. Differentiating instruction in response to student readiness, interest, and learning profile in academically diverse classrooms: A review of literature. *Journal for the Education of the Gifted* 27:119–145.
- Tomlinson, Carol Ann., and Jane M. Jarvis. 2023. Differentiation: Making curriculum work for all students through responsive planning & instruction. In *Systems and models for developing programs for the gifted and talented* (pp. 599–628). Routledge.
- Toyama, Michiko., and Yoshitaka Yamazaki. 2021. Classroom interventions and foreign language anxiety: A systematic review with narrative approach. In *Frontiers in Psychology* (Vol. 12). Frontiers Media S.A.
- Tran, Thi Trang Loan. 2022. An investigation into the causes of students' anxiety in learning English speaking skills. *International Journal of TESOL & Education* 2:183–196.
- Zarei, Abbas Ali., and Hossein Rezadoust. 2020. The effects of scaffolded and unscaffolded feedback on speaking anxiety and self-efficacy. *Journal of Modern Research in English Language Studies* 7:111–132.
- Zarin, Saima. 2025. A thematic review of the effects of speaking anxiety-reducing interventions on L2 speech producing efficiency in tertiary level. *Bulletin of Advanced English Studies* (BAES) 10.

Adriani Yulia Purwaningrum
 Faculty of Languages, Arts, and Culture
 Universitas Negeri Yogyakarta
 D.I. Yogyakarta, 55281, Indonesia
 adrianiyulipurwaningrum@uny.ac.id

基於分層教學的口語活動對大學一年級學生英語口語表現的影響

Adriani Yulia Purwaningrum

印尼日惹國立大學

本研究探討分層教學(Differentiated Instruction, DI)對英語作為外語(EFL)環境中大學一年級學生英語口語能力之影響。40名學生參與課堂行動研究(CAR)，採混合研究法收集資料：定量資料來自診斷測試、階段測試與後測，定性資料來自課堂觀察與學生反思。口語表現依CEFR標準，從流利度、準確性、互動、連貫性及表達能力評價。結果顯示，分層教學顯著提升學生英語口語表現與課堂參與度，並緩解口語焦慮。學生對分層口語活動普遍持積極態度，認為更符合個人能力與學習需求。本研究顯示，結合鷹架支持的分層教學能促進多元高等教育課堂中的英語口語發展，為多語言多元文化EFL教學提供實踐參考。

關鍵字：分層教學、英語作為外語(EFL)、外語焦慮、口語能力

INTEGRATING COLLABORATIVE ACTION RESEARCH, PROBLEM-BASED LEARNING, AND TEAM-BASED LEARNING IN A CROSS-NATIONAL CLIL CONTEXT: A CASE STUDY IN INTERNATIONAL MARKETING MANAGEMENT

Chih-Cheng Lin
Shih Chien University

ABSTRACT

This study examines how integrating Problem-Based Learning (PBL), Team-Based Learning (TBL), and Collaborative Action Research (CAR) within a Content and Language Integrated Learning (CLIL) environment is associated with student learning in an English-medium (EMI/CLIL) business course. Situated in Taiwan's Bilingual Nation by 2030 policy, it addresses the mismatch between exam-oriented business education and the communicative and intercultural competencies required in international workplaces. Drawing on Coyle's 4C framework, the study examines associations between an integrated CAR-PBL-TBL design and the four dimensions of Content, Cognition, Communication, and Culture, and between these dimensions and academic performance. The instructional design was implemented in an International Marketing Management course through cross-national collaboration with Green River College, USA. Data from 81 Taiwanese and American students were analyzed using PLS-SEM. Students reported significant enhancements across all four 4C dimensions. However, only content understanding and communication ability were significantly associated with semester grades; cognitive ability and cross-cultural understanding were not. The findings suggest that conventional grade-based assessments may not fully capture broader learning competencies in English-medium business education.

Keywords: Problem-Based Learning; Team-Based Learning; Collaborative Action Research; Content and Language Integrated Learning; bilingual education; instructional practice

1. INTRODUCTION

In an era of accelerating globalization and knowledge-intensive competition, language proficiency and intercultural competence have become widely recognized as foundational to international business success (Bartlett and Ghoshal 1998; Hofstede 2001). Taiwan's ambition to achieve full bilingualism by 2030 has placed considerable pressure on higher education institutions (particularly business schools) to restructure their pedagogical practices accordingly. Yet conventional assessment models in business programs, which tend to privilege examination performance over communicative and intercultural capability, remain poorly aligned with the competency profiles that industry requires (Knight and de Wit 1997). This misalignment raises pressing questions about how bilingual business education can be redesigned to cultivate both disciplinary expertise and real-world language use.

The urgency of this challenge is reinforced by the evolving demands of the global labor market. Multinational corporations increasingly seek graduates who are not only technically competent in their disciplines but also capable of negotiating meaning across cultural boundaries, facilitating collaborative decision-making in multilingual environments, and adapting their communicative strategies to diverse international audiences. Business schools in Taiwan, like those across much of East Asia, have historically concentrated on transmitting disciplinary knowledge through lecture-centered pedagogies and high-stakes written examinations. While this approach produces students with solid factual foundations, it tends to underinvest in the communicative fluency, intercultural sensitivity, and collaborative reasoning skills that the contemporary workplace demands. Taiwan's bilingual policy, rather than representing a purely linguistic ambition, thus reflects a deeper recognition that the country's economic competitiveness depends on

producing graduates who can operate authentically and effectively in international professional contexts.

The present study responds to this challenge by integrating three pedagogically compatible approaches within a fully English-medium instruction (EMI) business course: Problem-Based Learning (PBL), Team-Based Learning (TBL), and Collaborative Action Research (CAR). PBL places authentic, ill-structured problems at the center of the learning experience, motivating students to construct knowledge through guided inquiry and peer collaboration (Savin-Baden and Major 2004; Schmidt et al. 2011). TBL extends collaborative learning through structured team accountability mechanisms (including individual and team readiness assurance tests) that promote both individual preparation and collective problem-solving (Michaelsen and Sweet 2008). CAR frames the instructor as a reflective practitioner who systematically observes, documents, and refines instructional decisions through iterative cycles of planning, action, and reflection (Carr and Kemmis 1986). While each approach has an independent evidence base, their combined potential in professional business contexts remains underexplored.

What distinguishes the present integration from prior applications of these individual approaches is the deliberate architectural alignment among them. CAR does not merely accompany PBL and TBL as an evaluation strategy; it provides the regulatory infrastructure through which the other two approaches are continuously calibrated. This means that instructional choices are not fixed at the outset of the course but are treated as hypotheses to be tested, refined, and revised on the basis of empirical evidence gathered from students' responses, team interactions, and performance trajectories. The result is a pedagogical system that is simultaneously structured enough to produce reliable learning conditions and flexible enough to respond adaptively to the specific challenges that emerge in a cross-national bilingual context.

This integrated design is situated within a CLIL framework (Coyle et al. 2010; Dalton-Puffer 2007), which positions subject-matter learning and second-language acquisition as mutually reinforcing rather than competing objectives. CLIL's 4C framework (Content, Cognition, Communication, and Culture) provides a theoretically grounded set of mediating constructs through which the study examines how the CAR–

PBL–TBL instructional design shapes student learning. The intervention was operationalized in an International Marketing Management course, with a cross-national collaboration with a U.S. partner institution providing an authentic bilingual and intercultural context.

Three research objectives guide the study: (1) to develop and validate an integrated CAR–PBL–TBL instructional framework applicable to business school curricula; (2) to examine this framework’s association with student learning within a CLIL context aligned with Taiwan’s bilingual education policy; and (3) to quantitatively assess student learning outcomes under this framework using PLS-SEM. The central research question is: How is an integrated CAR–PBL–TBL instructional design, mediated by the four 4C dimensions of CLIL, associated with students’ 4C development and learning performance in an international marketing course?

2. THEORETICAL BACKGROUND AND HYPOTHESIS DEVELOPMENT

2.1 Collaborative Action Research

Collaborative Action Research (CAR) is a participatory, practice-oriented methodology in which teachers engage in structured cycles of planning, action, observation, and reflection to interrogate and improve their own instructional practice (Carr and Kemmis 1986). Its epistemological roots lie in the reflective traditions of Locke and Dewey: Locke’s contention that all knowledge originates in sensation and reflection, and Dewey’s pragmatist insistence that genuine learning requires both active engagement and deliberate self-examination. These philosophical antecedents converge with more recent constructivist accounts of professional knowledge, which hold that cognitive structures can only be meaningfully reorganized through reflective activity, an insight that Polya (1957) operationalized in the context of mathematical problem-solving and that remains relevant to pedagogical inquiry.

Within educational research, the significance of teacher reflection has been extensively documented. The Cognitively Guided Instruction program developed at the University of Wisconsin demonstrated that teachers' reflective capacity is at least as consequential as their content knowledge for shaping student learning (Fennema and Franke 1992). Brown and Borko (1992) similarly identified reflective competence as a constitutive element of effective teaching. Noffke and Zeichner (1987) found that sustained reflective practice reshapes teachers' professional identities, sharpens instructional decision-making, and fosters greater criticality and creativity in classroom contexts. Reflection, in this body of work, is not merely an add-on to teaching but a catalyst for the reconstruction of practice-relevant knowledge.

A further consideration is that the reflective cycle in CAR is not a singular event but an iterative spiral that becomes progressively more refined across the duration of an intervention. Each planning–action–observation–reflection loop generates evidence that informs the next loop, enabling the instructor to move from general instructional intentions to increasingly precise, evidence-grounded responses to specific student needs. This iterative quality makes CAR especially well suited to complex pedagogical environments, such as cross-national bilingual courses, where the instructional challenges that arise cannot be fully anticipated at the outset and where generic, pre-specified lesson plans are unlikely to remain adequate across the full arc of a semester. In this sense, CAR does not merely evaluate a fixed pedagogical design; it constitutes a mechanism through which the design continuously evolves.

The collaborative dimension of CAR amplifies these individual benefits by distributing inquiry across a professional community. Research teams (which may include classroom teachers, administrators, and university-based researchers) follow the same spiral of inquiry while pooling observations, challenging interpretations, and generating shared knowledge (Banegas et al. 2013; Timperley et al. 2014). In the present study, this collaborative structure was instantiated through a partnership between the course instructor, a faculty colleague from Green River College (USA), and a research assistant. During the planning phase, the team co-designed the curriculum around PBL and TBL tasks. During each action phase, the instructor taught while the research assistant documented

classroom events through field notes and video recording. Reflection sessions following each observation allowed the team to review evidence, negotiate interpretations, and revise subsequent lesson plans. This iterative architecture served both a pedagogical function (ensuring the instructional design remained responsive to emerging student needs) and a methodological function, generating the systematic evidence needed to evaluate the intervention.

2.2 Problem-Based Learning

Problem-Based Learning (PBL) is a well-established constructivist pedagogy that organizes learning around authentic, complex problems whose resolution requires students to actively seek, evaluate, and integrate new knowledge (Hmelo-Silver 2004; Savery 2006). Foundational accounts characterize PBL as student-centered and self-directed: rather than receiving transmitted content, students diagnose problems, formulate learning issues, identify and consult relevant resources, and collaboratively construct solutions, with the instructor functioning as a facilitator rather than an information source (Savin-Baden and Major 2004; Taylor and Mifflin 2008). This orientation aligns with McKeachie and Svinicki's (2006) argument that active problem engagement produces deeper and more transferable understanding than passive reception.

Small-group deliberation is central to the PBL process. Peer interaction exposes students to multiple interpretive frameworks, creates productive cognitive conflict, and necessitates the kind of verbal articulation of reasoning that consolidates conceptual understanding (Schmidt et al. 2011). Research applying PBL in higher education has consistently documented positive effects on motivation, deep learning, and self-regulated inquiry (Dolmans et al. 2016; Hmelo-Silver et al. 2007). In Taiwanese higher education contexts, studies have confirmed PBL's effectiveness in stimulating intrinsic motivation and improving academic outcomes in both quantitative and social science disciplines (Ho 2017).

A further strength of PBL lies in its capacity to develop transferable competencies that extend beyond the specific problem being addressed. When students repeatedly practise the diagnostic cycle of problem identification, information search, hypothesis formation, and evaluative

synthesis, they internalize a generalizable inquiry schema that can be applied to novel challenges across professional contexts. This transfer potential is particularly valuable in business education, where the complexity and unpredictability of real organizational environments make the capacity for adaptive problem-solving more consequential than the ability to reproduce a fixed body of knowledge. PBL thus operates not only as a vehicle for disciplinary learning but as a scaffold for professional identity development, cultivating the disposition to approach ill-structured challenges with structured curiosity rather than seeking algorithmic solutions.

In the present International Marketing Management course, PBL was deployed to engage students with real-world market entry challenges. Working in small groups, students assumed the role of management consultants advising companies on cross-border marketing strategy. Taiwanese student teams developed recommendations for a brand's entry into the U.S. market, while their American counterparts reciprocally developed plans for the Taiwanese market. This design not only situated learning in professionally authentic contexts but also created a natural scaffold for cross-cultural perspective-taking, an alignment with CLIL's Culture dimension that will be further theorized below.

2.3 Team-Based Learning

Team-Based Learning (TBL) is a structured form of collaborative pedagogy comprising three sequential phases: a pre-class preparation stage, a readiness assurance process, and a team application stage (Michaelson and Sweet 2008). In the preparation stage, students study assigned materials independently before class and are organized into stable, heterogeneous groups of five to eight members. The readiness assurance process begins with an Individual Readiness Assurance Test (iRAT), followed by the same test administered to teams (tRAT), and concludes with team-level appeals and instructor clarification. The application stage presents teams with complex, discipline-specific problems that require integrating and extending the pre-class material, tasks designed to develop higher-order reasoning rather than mere recall.

Michaelsen et al. (2008) identify group composition, individual accountability, timely feedback, and assignment design as the four structural pillars of TBL. When these elements are properly implemented, TBL reliably produces learning teams rather than merely assigned groups, collectives capable of generating intellectual outputs that no individual member could produce alone. The accountability mechanisms embedded in TBL are particularly important: by requiring both individual preparation and collective performance, TBL creates conditions under which students are genuinely motivated to arrive prepared and to contribute substantively to team deliberations. Research in medical education (Parmelee et al. 2009) and more broadly in higher education (Haidet et al. 2012) confirms TBL's capacity to improve knowledge retention, enhance communication skills, and foster a sense of shared responsibility for learning outcomes. In the context of the present study, TBL's emphasis on team communication and collective accountability directly targets the Communication and Cognition dimensions of the 4C framework.

The iRAT–tRAT sequence deserves particular attention as a pedagogical mechanism. The individual test ensures that each student arrives with a baseline of content knowledge, preventing the free-rider dynamics that often undermine group work when preparation is voluntary and unverified. The team test then transforms this individual preparation into collective deliberation, as team members must negotiate disagreements, articulate reasoning, and reach consensus under time pressure. This transition from individual to collective sense-making is itself a form of active learning, one that replicates the conditions of professional team decision-making far more faithfully than individual written examinations. The instructor clarification phase that follows provides immediate, corrective feedback precisely when student motivation to understand their errors is highest, reinforcing the principle, articulated by Hattie and Timperley (2007), that feedback is most effective when it is timely, specific, and embedded in an active learning context.

2.4 Content and Language Integrated Learning

Content and Language Integrated Learning (CLIL) was first systematized by David Marsh in 1994 (Coyle et al. 2010) as a dual-focused educational approach in which subject matter is taught through the medium of an additional language. Its foundational premise is that meaningful language acquisition occurs most efficiently when learners are cognitively engaged with substantive content, a position consistent with both Krashen's comprehensible input hypothesis and more recent usage-based theories of language development (Mehisto et al. 2008). Rather than treating language as an object of study, CLIL positions it as a vehicle for disciplinary knowledge construction, creating conditions under which vocabulary, syntax, and pragmatic competence develop as organic by-products of subject engagement.

Coyle's (2005) 4C framework articulates the four dimensions along which CLIL instruction operates. Content refers to the disciplinary subject matter students engage with, encompassing both canonical and interdisciplinary knowledge. Cognition denotes the higher-order thinking operations (comparison, analysis, synthesis, evaluation) through which students process and internalize that content. Communication encompasses not merely the target language itself but the full range of discourse practices through which subject knowledge is constructed, negotiated, and expressed. Culture, finally, refers to the intercultural awareness and sensitivity that naturally develops when learners engage with disciplinary perspectives and interlocutors shaped by different cultural formations. Together, these four dimensions provide a principled basis for evaluating whether a CLIL intervention has achieved its dual linguistic and disciplinary objectives.

A distinctive feature of CLIL is the productive tension it introduces between linguistic and cognitive demands. When learners must simultaneously manage the cognitive load of grappling with new disciplinary concepts and the linguistic challenge of encoding and decoding those concepts in an additional language, they are compelled to attend more carefully to both dimensions than they might in either a monolingual content class or a language class focused on simplified communicative tasks. This dual-demand structure, when scaffolded

appropriately, promotes the kind of deep processing that produces durable learning. Empirical evaluations of CLIL in higher education contexts have documented gains across all four dimensions, though effects vary by context, learner population, and implementation quality (Dalton-Puffer et al. 2010). Particularly relevant to the present study are findings demonstrating that CLIL environments foster intercultural communication competence when students must actively negotiate meaning across cultural boundaries, precisely the conditions created by the cross-national collaboration with the U.S. partner institution.

The alignment between CLIL and the CAR–PBL–TBL integration is architecturally coherent rather than incidental. PBL’s authentic, content-rich problems provide the disciplinary substance that CLIL requires learners to engage with in the target language. TBL’s structured team accountability ensures that this engagement is not passive but involves the kind of interactive, negotiated language use that CLIL theory identifies as the engine of communicative development. And CAR’s reflective cycles allow the instructor to monitor the quality of this engagement and to make targeted adjustments when the dual demands of content and language overwhelm students’ current competence levels. The 4C framework thus functions not only as an analytical lens for evaluating the intervention but as a design template that guided its construction.

2.5 Hypothesis Development

The conceptual framework of this study positions the integrated CAR–PBL–TBL instructional design as an independent construct whose effects on learning performance are mediated by the four 4C dimensions. The rationale for this mediation structure is as follows: effective CAR–PBL–TBL design provides students with the structural conditions (authentic problems, team accountability, iterative instructor feedback) needed to engage deeply with disciplinary content (Content), develop higher-order reasoning (Cognition), practice collaborative discourse (Communication), and navigate intercultural differences (Culture). These four capabilities, when cultivated, then translate into demonstrable learning performance. The following hypotheses formalize this logic.

Regarding content understanding: In PBL environments, students develop richer and more durable knowledge representations by actively grappling with real-world problems rather than passively receiving presented information (Hmelo-Silver 2004). TBL's readiness assurance mechanisms ensure that foundational content is reliably encoded before higher-order application tasks begin, while the CAR reflection cycle allows the instructor to identify and address persistent misconceptions (Dolmans et al. 2016; Parmelee et al. 2009; Savin-Baden 2014). Taken together, these mechanisms create multiple, reinforcing pathways for deep content engagement.

H1: An integrated CAR-PBL-TBL instructional design is positively associated with students' content understanding.

Regarding cognitive ability: PBL's problem-centered structure requires students to apply analysis, synthesis, and evaluative reasoning to ill-structured real-world challenges (Savery 2006). TBL's application phase (in which teams must defend positions and respond to peer challenges) further exercises higher-order cognition. The combination stimulates metacognitive awareness and the kind of flexible, transferable thinking that characterizes expert reasoning (Hmelo-Silver 2004; Michaelsen et al. 2004).

H2: An integrated CAR-PBL-TBL instructional design is positively associated with students' cognitive ability.

Regarding communication ability: TBL's team accountability structure makes interpersonal communication instrumentally necessary for learning success: students who fail to communicate effectively cannot fulfill their team responsibilities (Haidet et al. 2012; Michaelsen et al. 2004). PBL's small-group problem-solving contexts provide repeated, low-stakes opportunities for students to practice articulating complex ideas, listening actively, and negotiating disagreement. CAR's reflective observation function enables the instructor to provide targeted feedback on communicative effectiveness.

H3: An integrated CAR-PBL-TBL instructional design is positively associated with students' team communication ability.

Regarding cross-cultural understanding: The cross-national collaboration at the heart of this course creates authentic intercultural encounters that are rarely achievable in single-institution settings.

Working alongside peers from a different cultural background (and advising them on market entry into one's own cultural context) heightens awareness of cultural difference and develops the perspective-taking skills central to intercultural competence (Kember 2000). CLIL's explicit attention to the Culture dimension further scaffolds this development by encouraging reflection on how cultural assumptions shape disciplinary knowledge and professional practice.

H4: An integrated CAR-PBL-TBL instructional design is positively associated with students' cross-cultural understanding.

Regarding learning performance: A substantial body of evidence links each of the four 4C dimensions to academic outcomes. Students with stronger content understanding predictably demonstrate higher performance on knowledge assessments (Hmelo-Silver et al. 2007). Cognitive ability (particularly higher-order reasoning) predicts academic success across disciplines (Mevarech and Kramarski 1997). Effective team communication is positively associated with academic achievement in collaborative course formats (Blasco-Arcas et al. 2013; Springer et al. 1999). Cross-cultural competence has been linked to enhanced academic engagement and performance in internationalized learning environments (Dalton-Puffer 2011; Deardorff 2006). The present study operationalizes learning performance through semester grades independently assigned by instructors at each institution, and hypothesizes positive relationships with all four 4C mediators.

H5: Students' content understanding is positively associated with learning performance.

H6: Students' cognitive ability is positively associated with learning performance.

H7: Students' team communication ability is positively associated with learning performance.

H8: Students' cross-cultural understanding is positively associated with learning performance.

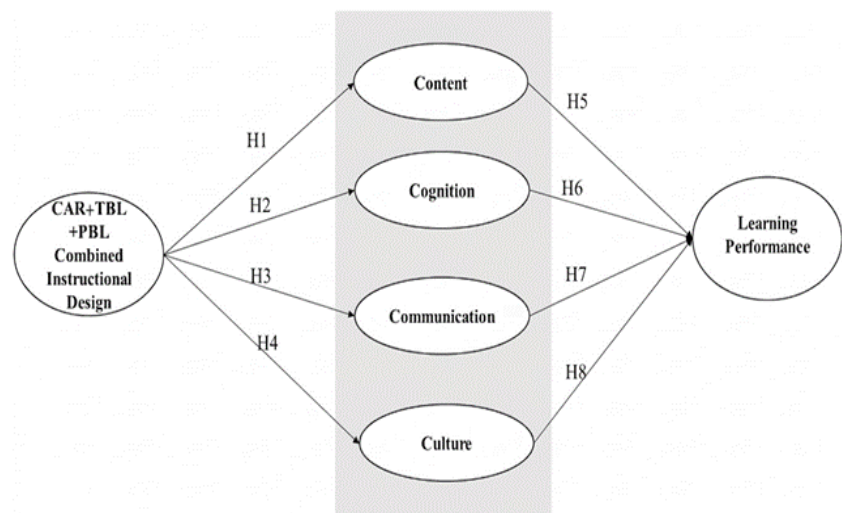


Figure 1 Conceptual Research Framework

3. RESEARCH DESIGN AND METHODOLOGY

3.1 Course Setting and Instructional Implementation

The empirical context for this study is an International Marketing Management course delivered at a Taiwanese university, with enrolled students serving as the primary participant group. A faculty member and students from Green River College (Washington State, USA) participated as cross-national collaborative partners. This pairing was designed to create an authentic bilingual and intercultural learning environment consistent with CLIL principles, one in which content learning and second-language use were genuinely interdependent rather than artificially juxtaposed.

The selection of Green River College as the partner institution was motivated by practical as well as pedagogical considerations. Practically,

the institution's geographic and temporal proximity to major U.S. consumer markets made its students credible sources of local market intelligence for the Taiwanese teams' consultancy projects. Pedagogically, the community college context ensured that American participants brought varied disciplinary backgrounds and professional experiences to the collaboration, enriching the diversity of perspectives available for cross-cultural comparison. The faculty coordinator at Green River College participated as both a teaching collaborator and a co-investigator within the CAR team, contributing observations of student behavior from the American side of the collaboration and thereby ensuring that the reflective inquiry encompassed both national cohorts.

The instructional plan unfolded in three phases. During Phase 1 (Weeks 1–5), the instructor established the disciplinary foundations of international marketing through structured input, with readings, lectures, and concept-building activities designed to prepare students for the more complex collaborative tasks ahead. During Phase 2 (Weeks 6–14), PBL and TBL strategies came to the fore: student teams selected a real brand and, positioned as management consultants, developed a market entry strategy for their counterparts' home market. Taiwanese teams produced recommendations for U.S. market entry; American teams developed plans targeting the Taiwanese market. Peer review across institutions was built into this phase to ensure genuine intercultural exchange. During Phase 3 (Weeks 15–18), teams presented their strategic analyses in synchronous online sessions attended by both cohorts, enabling direct cross-cultural dialogue and instructor evaluation. In Week 18, post-intervention data collection was completed through questionnaire administration and final grade assignment by instructors from both institutions.

To clarify the EMI/CLIL implementation, the course was conducted as a fully English-taught International Marketing Management course. English served as the medium for readings, lectures, cross-national peer interaction, written project outputs, and the Phase 3 synchronous presentations. Chinese was not used as a regular instructional language; it appeared only minimally and informally when Taiwanese students needed brief within-group clarification or administrative confirmation before returning to English for class participation and project work. Thus, the

course should be understood as a full EMI/CLIL course with very limited first-language support rather than as a balanced two-language course.

The synchronous online presentation format adopted in Phase 3 merits particular attention as a CLIL mechanism. Unlike asynchronous interactions (such as written peer reviews or discussion board posts), synchronous video-mediated communication imposes real-time communicative demands that more closely approximate professional international business meetings. Students could not revise or polish their language in advance; they were required to listen, interpret, and respond to interlocutors across a cultural and linguistic boundary in real time. This format created the kind of authentic communicative pressure that CLIL theorists identify as particularly generative for language and intercultural development, and it provided the CAR team with rich observational data about students' communicative competencies under realistic conditions.

Throughout all three phases, the CAR cycle operated in parallel: the instructor maintained a reflective journal documenting observations of student progress and instructional decisions; the research assistant video-recorded sessions and prepared structured observation protocols; and regular research meetings between the two faculty collaborators and the research assistant reviewed this evidence and generated revisions to subsequent lesson plans.

3.2 Participants

Eighty-one students participated in the study: 41 enrolled at the Taiwanese institution and 40 at Green River College. The Taiwanese cohort was composed predominantly of graduate students in a business administration program, the majority of whom were completing their second or third year of postgraduate study. The American cohort drew from undergraduate students enrolled in business and international studies programs. This asymmetry in academic level was acknowledged in the CAR team's planning deliberations and addressed through scaffolding adjustments in Phase 1, which provided American students with additional background on Taiwanese business culture and provided Taiwanese

students with orientation materials on the American retail and consumer landscape.

Regarding the Taiwanese students' English profile, formal standardized CEFR scores were not collected as part of the research protocol. However, the Taiwanese program requires students to meet an English graduation threshold of TOEIC 820 or its institutionally recognized equivalent, a level that, according to widely used concordance tables, corresponds approximately to the upper B2 to C1 range of the Common European Framework of Reference (CEFR). This threshold provides an institutionally defined proficiency benchmark for students in the English-taught curriculum. In addition, because the Taiwanese cohort was enrolled in a fully English-medium business program and had completed prior English-taught business coursework, their functional academic English use can be reasonably characterized as high-intermediate to advanced for course participation. This TOEIC-based benchmark and its approximate CEFR mapping are reported only as contextual evidence of the cohort's expected academic English profile and should not be interpreted as individually measured proficiency variables in the structural model.

The cross-national composition of the sample is both a strength and a design consideration of the study. On the one hand, the inclusion of genuine international partners, rather than simulated cross-cultural interactions, creates the authentic intercultural conditions that the CLIL framework requires and that most single-institution studies cannot replicate. On the other hand, cross-national samples introduce heterogeneity in educational backgrounds, language proficiency levels, and familiarity with collaborative pedagogies that must be acknowledged when interpreting findings. The research design sought to address this by maintaining parallel instructional structures across both cohorts, using the same questionnaire instruments (with a Chinese translation provided for Taiwanese students), and ensuring that grading rubrics were discussed and aligned between the two instructors before the course began.

3.3 Measures

Data collection comprised two components: course grades and a self-report questionnaire. Grades served as the outcome variable and were independently assigned by instructors from both institutions; Taiwanese grades were recorded as raw percentage scores, while American grades were converted to a comparable percentage scale.

To improve comparability across the two institutional contexts, the two instructors discussed the assessment scheme before the course began and aligned grading around a shared set of performance criteria: (a) accuracy and depth of international marketing content, (b) quality of market evidence and strategic reasoning, (c) clarity and persuasiveness of English communication in written and oral outputs, (d) responsiveness to cross-national peer feedback, and (e) contribution to team-based project work. Within each course, these criteria were operationalized through three principal graded components whose relative emphasis was kept broadly comparable: the final market-entry project report (weighted most heavily, in the region of 40 percent of the course grade), the Phase 3 synchronous cross-national presentation (approximately 30 percent), and ongoing team contribution and peer-feedback responsiveness (approximately 30 percent). Because the two institutions used different internal grade categories and conversion conventions, the study did not assume identical weighting across sites; instead, final percentage grades were compared only after both instructors mapped these major components onto the shared performance criteria, with the final market-entry project and synchronous presentation serving as the most directly comparable components.

The questionnaire assessed five constructs. The integrated CAR–PBL–TBL design was measured with five items adapted from Mayer and Johnson (2010), tapping students' evaluations of the approach as an effective, impactful, and satisfying learning tool. Content understanding was assessed with four items (Paas 1992) capturing students' ability to summarize, explain, and apply course concepts. Cognitive ability was measured with six items drawing on Schraw and Dennison (1994) and Pintrich and De Groot (1990), covering analytical reasoning, complex problem-solving, and task concentration. Communication ability was

assessed with four items (Chang et al. 2006; Ellis and Goodyear 2010) addressing oral expression, interpersonal clarity, listening competence, and comfort in group discussion. Cross-cultural understanding was measured with four items (Chen and Starosta 2000; Deardorff 2006) addressing appreciation of cultural difference, knowledge of foreign customs, comfort in intercultural interaction, and adaptive flexibility. All items used a seven-point Likert response format (1 = strongly disagree, 7 = strongly agree). Detailed item wording is provided in the Appendix.

The use of established, previously validated scales represents a deliberate methodological choice. Rather than developing new instruments from scratch, the study adapts items from well-cited educational psychology and intercultural communication literatures, thereby benefiting from prior evidence of construct validity while tailoring the item wording to the specific CAR–PBL–TBL context. Each adapted item was reviewed by the CAR team during the planning phase and refined to ensure linguistic appropriateness for both English-speaking and Chinese-speaking respondents. The Chinese translation of the questionnaire was independently back-translated by a bilingual research collaborator to verify semantic equivalence.

In addition to questionnaire and grade data, the CAR process generated complementary observational and performance-based evidence, including instructor reflective journals, research-assistant field notes, video records of selected sessions, and Phase 3 presentation observations. These materials were used to triangulate the interpretation of the quantitative findings and to guide iterative instructional refinements. They were not entered as additional indicators in the PLS-SEM model because the present analysis was designed to test the hypothesized latent-variable structure using standardized survey and grade data; nevertheless, they provide contextual support for the reported patterns, especially the strong role of communication ability.

3.4 Analytical Approach

The study employed partial least squares structural equation modeling (PLS-SEM), implemented in SmartPLS, as the primary analytical method.

PLS-SEM was selected for four reasons. First, it accommodates the relatively small combined sample ($N = 81$) more effectively than covariance-based SEM methods such as maximum likelihood estimation, which require substantially larger samples to produce stable estimates (Hair et al. 2017). Second, PLS-SEM is designed for prediction-oriented research with complex models containing multiple latent variables, a design feature that matches the present mediation model (Hair et al. 2011). Third, it handles the multicollinearity that is common among theoretically related constructs in educational research (Chin 2010; Henseler et al. 2009). Fourth, it permits the inclusion of single-indicator constructs alongside multi-item scales, enabling the direct integration of the grade-based performance variable into the structural model (Hulland 1999). Analysis proceeded in two stages: assessment of the outer (measurement) model for reliability and validity, followed by testing of the inner (structural) model path coefficients.

Bootstrapping with 5,000 resamples was used to generate confidence intervals and t-statistics for all path coefficients, following established recommendations for PLS-SEM inference (Hair et al. 2017). Two-tailed significance tests were applied throughout, with $\alpha = .05$ as the threshold for statistical significance. Effect sizes were evaluated using the coefficient of determination (R^2) for endogenous constructs and the standardized path coefficient (β) for individual hypothesized relationships. The assessment of discriminant validity followed the Fornell-Larcker criterion, under which the square root of each construct's average variance extracted (AVE) must exceed its correlations with all other constructs. Multicollinearity among predictor constructs was evaluated using variance inflation factors (VIF), with values below 5.0 indicating acceptable levels.

4. RESULTS

4.1 Measurement Model Assessment

Table 1 presents the reliability and validity statistics for all constructs. Internal consistency was evaluated using Cronbach's α and composite reliability (CR). Cronbach's α values ranged from .92 to .96, comfortably exceeding Nunnally's (1978) recommended threshold of .70. CR values ranged from .93 to .97, surpassing the .60 criterion proposed by Fornell and Larcker (1981). Average variance extracted (AVE) values ranged from .78 to .88, all exceeding the .50 threshold that indicates convergent validity (Fornell and Larcker 1981). Discriminant validity was evaluated using the Fornell-Larcker criterion: the square root of each construct's AVE (shown on the diagonal of Table 1) exceeded all inter-construct correlations in the corresponding row and column, confirming adequate discriminant validity.

Table 1 Construct Reliability, Validity, and Intercorrelations

Construct	1	2	3	4	5	6
1. CAR-PBL-TBL Integration	.886					
2. Content Understanding	.878	.937				
3. Cognitive Ability	.792	.695	.927			
4. Communication Ability	.818	.817	.676	.900		
5. Cross-Cultural Understanding	.810	.808	.689	.744	.933	
6. Learning Performance	.867	.871	.781	.805	.812	1.000
CR	.967	.949	.945	.955	.928	
AVE	.870	.860	.784	.879	.810	
Cronbach's α	.963	.946	.945	.954	.922	

Note. Bold diagonal values = $\sqrt{\text{AVE}}$. N = 81. CAR = Collaborative Action Research; PBL = Problem-Based Learning; TBL = Team-Based Learning. CR = composite reliability; AVE = average variance extracted.

Factor loadings for individual items are reported in Table 2. All loadings ranged from .84 to .96, surpassing Hair et al.'s (1998) recommended minimum of .50. Variance inflation factor (VIF) values for all indicators were below 5.0 (range: 2.50–4.97), indicating that

multicollinearity does not threaten the integrity of the structural estimates (Hair et al. 2011). The coefficient of determination (R^2) for the endogenous constructs ranged from .47 to .83, reflecting moderate to substantial explanatory power across the model.

The uniformly high factor loadings across all constructs provide strong evidence that the items are measuring their intended constructs with precision. Of particular note is the consistently high loading pattern for the CAR–PBL–TBL integration construct (loadings: .897–.959), which suggests that students formed a coherent and internally consistent impression of the integrated instructional design as a unified learning experience rather than perceiving its three components (CAR, PBL, and TBL) as disconnected elements. This finding supports the theoretical premise of the study, which treats the three-approach integration as a single, unified construct rather than a mere aggregation of independent pedagogies.

Table 2 Factor Loadings and Variance Inflation Factors for Construct Indicators

Indicator	Factor Loading	VIF
CAR–PBL–TBL Integration 1	.959	4.08
CAR–PBL–TBL Integration 2	.897	3.47
CAR–PBL–TBL Integration 3	.941	3.55
CAR–PBL–TBL Integration 4	.921	3.85
CAR–PBL–TBL Integration 5	.945	4.34
Content Understanding 1	.911	3.86
Content Understanding 2	.921	4.60
Content Understanding 3	.944	3.14
Content Understanding 4	.933	4.06
Cognitive Ability 1	.900	4.97
Cognitive Ability 2	.933	4.44
Cognitive Ability 3	.884	3.31
Cognitive Ability 4	.854	2.67
Cognitive Ability 5	.860	2.95

Indicator	Factor Loading	VIF
Cognitive Ability 6	.880	3.68
Communication Ability 1	.952	3.13
Communication Ability 2	.927	3.45
Communication Ability 3	.929	3.42
Communication Ability 4	.941	3.86
Cross-Cultural Understanding 1	.918	3.15
Cross-Cultural Understanding 2	.912	3.65
Cross-Cultural Understanding 3	.925	3.96
Cross-Cultural Understanding 4	.843	2.50

Note. All loadings exceeded the .50 threshold recommended by Hair et al. (1998). All VIF values < 5.0.

4.2 Structural Model and Hypothesis Testing

Table 3 reports the path coefficients, standard errors, t-values, p-values, and decisional outcomes for each hypothesized relationship. H1 through H4 examined the direct paths from the CAR–PBL–TBL design to the four 4C dimensions. All four paths reached significance at $p < .001$. The integrated instructional design showed the strongest association with cognitive ability ($\beta = .810$, $t = 27.63$) and communication ability ($\beta = .808$, $t = 21.39$), followed by cross-cultural understanding ($\beta = .744$, $t = 11.93$) and content understanding ($\beta = .689$, $t = 10.90$). H1 through H4 were therefore all supported.

H5 through H8 examined the relationships between the four 4C dimensions and learning performance. Content understanding ($\beta = .242$, $t = 3.63$, $p < .001$) and communication ability ($\beta = .415$, $t = 3.81$, $p < .001$) both significantly predicted semester grades; H5 and H7 were supported. By contrast, cognitive ability ($\beta = .193$, $t = 1.29$, $p = .198$) and cross-cultural understanding ($\beta = .144$, $t = 1.48$, $p = .140$) did not reach significance; H6 and H8 were not supported.

Table 3 PLS-SEM Path Coefficients for the Structural Model

Path	β	SE	t	p	Decision
H1: CAR–PBL–TBL → Content Understanding	.689	.063	10.90	< .001	Supported
H2: CAR–PBL–TBL → Cognitive Ability	.810	.029	27.63	< .001	Supported
H3: CAR–PBL–TBL → Communication Ability	.808	.038	21.39	< .001	Supported
H4: CAR–PBL–TBL → Cross-Cultural Understanding	.744	.062	11.93	< .001	Supported
H5: Content Understanding → Learning Performance	.242	.067	3.63	< .001	Supported
H6: Cognitive Ability → Learning Performance	.193	.150	1.29	.198	Not supported
H7: Communication Ability → Learning Performance	.415	.109	3.81	< .001	Supported
H8: Cross-Cultural Understanding → Learning Performance	.144	.098	1.48	.140	Not supported

Note. β = standardized path coefficient; SE = bootstrap standard error (5,000 resamples); t = t-statistic; p = two-tailed p-value.

5. DISCUSSION

5.1 Summary of Findings

The results collectively support the central premise of this study: an integrated CAR–PBL–TBL instructional design, implemented within a CLIL environment, was associated with stronger reported development in the four competency dimensions articulated by Coyle's (2005) 4C framework. The particularly strong associations with cognitive ability and communication ability (both $\beta \approx .81$) are theoretically coherent: PBL and TBL are most directly designed to provoke higher-order reasoning and to demand interpersonal communication, respectively. The somewhat weaker (though still significant) association with content understanding ($\beta = .69$) likely reflects the inherently greater variability in students' prior

disciplinary knowledge compared with their reasoning and communication experiences.

A more notable finding concerns the differentiated pattern of associations with learning performance. Content understanding and communication ability were both positively associated with grades; cognitive ability and cross-cultural understanding were not. One interpretation is that the semester grade (independently assigned by instructors at each institution) remains primarily sensitive to knowledge recall and interpersonal presentational competence, dimensions that map directly onto written and oral assessments. Cognitive ability, however it develops within CLIL–PBL–TBL contexts, may manifest in forms of reasoning (nuanced argumentation, epistemic flexibility, tolerance for ambiguity) that conventional assessment rubrics do not reliably capture. Cross-cultural understanding may similarly develop along trajectories that unfold over longer timescales than a single semester affords, and that require assessment instruments sensitive to dispositional change rather than knowledge gain.

Two explanations should be considered when interpreting the non-significant H6 and H8 paths. First, the grading system may have been more sensitive to observable content mastery and communicative performance than to deeper cognitive growth or intercultural disposition, which are less easily captured by end-of-semester course grades. Second, the modest sample size ($N = 81$), combined with the high intercorrelations among the 4C dimensions, may have reduced the model's ability to detect smaller unique effects once content understanding and communication ability were included simultaneously. The non-significant findings therefore should not be read as evidence that cognition or culture were unimportant; rather, they suggest that these dimensions may require more targeted assessment instruments and larger samples to estimate their independent contribution to performance.

It is also instructive to consider the relative magnitudes of the significant paths to learning performance. Communication ability ($\beta = .415$) showed a substantially stronger association with grades than content understanding ($\beta = .242$), suggesting that the assessment practices in this course, which included cross-national oral presentations and peer reviews, may have been particularly well calibrated to reward

communicative competence. This finding carries a design implication: when instructors deliberately build communicative assessment tasks into a CLIL course, they not only create opportunities for language development but also align the grading system more closely with the communicative outcomes that the pedagogy targets. The gap between the two significant paths thus reflects a partial success story: the course's assessment design captured the communication outcomes but did not fully instrument the deeper cognitive and intercultural gains that students were simultaneously developing.

The observational evidence gathered through the CAR process is consistent with this interpretation. During the Phase 3 synchronous presentations, instructor field notes and session recordings indicated that students frequently had to clarify market assumptions on the spot, respond to unscripted questions from peers based in the target market, and reformulate their explanations in real time when initial phrasings were not understood across the linguistic boundary. Observers noted that the teams whose oral performance showed the greatest spontaneous fluency and responsiveness under these conditions were generally those that also received the strongest assessment outcomes, whereas variation in written content accuracy was comparatively narrow. Although these observations were not entered into the PLS-SEM model, they offer concrete qualitative support for the quantitative result that communication ability was the strongest predictor of learning performance.

5.2 Theoretical Contributions

This study makes two theoretical contributions to the educational literature. First, it proposes the CAR–PBL–TBL integration as a novel instructional construct. While each of the three component approaches has an established theoretical and empirical profile, the present study is the first to conceptualize their systematic combination as a coherent pedagogical design framework and to subject that combination to empirical validation. The three approaches appear to reinforce one another in this instructional design: CAR provides the reflective infrastructure through which PBL and TBL are continuously refined; PBL and TBL

supply the active learning structures that CAR's reflective cycles have material evidence to evaluate.

The theoretical novelty of this integration lies in the structural complementarity of its three components. PBL and TBL both operate at the level of the learning task, defining how students engage with content and with one another. CAR operates at the level of the instructional system, defining how the instructor monitors and adjusts the learning environment in response to evidence. This vertical differentiation, between task-level and system-level pedagogical mechanisms, is what gives the CAR-PBL-TBL integration its adaptive capacity. An instructional design built on PBL and TBL alone would benefit from their intrinsic learning-activation properties but would lack a formal mechanism for systematic refinement across the course timeline. Adding CAR therefore converts PBL and TBL from a fixed instructional plan into a self-correcting system, one that can respond to emergent challenges, capitalize on unanticipated learning opportunities, and generate cumulative instructional wisdom that informs future iterations of the course.

Second, the study advances CLIL scholarship by demonstrating the 4C framework's utility as a set of mediating constructs linking instructional design to academic performance, a role it has not previously played in empirically tested structural models. The finding that Content and Communication mediate the path to grades while Cognition and Culture do not suggest that conventional course grades may capture only part of the learning outcomes targeted in bilingual business education. This insight calls for a more expansive conceptualization of learning outcomes in CLIL contexts, one that treats measurable academic performance as a partial and potentially misleading proxy for the richer competency development that CLIL interventions are designed to produce.

5.3 Practical Recommendations

Several practical implications follow ~~from~~ these findings. Educators designing bilingual business courses should consider adopting the integrated CAR-PBL-TBL framework as a structured approach to achieving CLIL's dual goals. Practically, this means building course

architectures around authentic, cross-culturally situated problems (PBL), organizing students into accountable, stable learning teams with clear individual and collective responsibilities (TBL), and maintaining a systematic reflective practice through which instructors document, analyze, and revise their instructional decisions (CAR).

For institutions implementing Taiwan's bilingual education policy, the cross-national collaboration model piloted in this study offers a replicable design for achieving authentic CLIL conditions at scale. Partnerships with English-medium institutions in other countries can be established through existing university exchange networks and formalized through memoranda of understanding that specify mutual teaching commitments and data-sharing protocols. The synchronous online presentation format used in this study is particularly cost-effective, as it achieves genuine intercultural interaction without the logistical and financial burdens of international travel. With appropriate facilitation, such partnerships can generate intercultural learning experiences that rival or exceed what students might encounter in short-term study-abroad programs, while allowing them to remain embedded in their home institution's degree structure.

The non-significant path from cognitive ability to learning performance suggests that current assessment practices may need recalibration. Project-based assessments, reflective portfolios, and oral defense formats are better positioned than time-limited written examinations to capture the analytical depth and argumentative sophistication that PBL and TBL environments cultivate. Similarly, the non-significant path from cross-cultural understanding to grades may indicate a need for longitudinal assessment designs (such as repeated cultural reflection tasks or pre-post intercultural competence inventories) that can track dispositional development over time rather than relying on end-of-semester self-report snapshots.

Instructors implementing CLIL in business contexts should also attend to the structural conditions that produced strong effects on Communication: namely, repeated, low-stakes opportunities for students to articulate disciplinary reasoning in English, receive feedback on that articulation, and refine their communicative practice in response. The cross-national collaboration design was particularly effective here, as it

created an authentic audience for students' communication that carried genuine social and intellectual stakes beyond the instructor evaluation. When students know that their market entry recommendations will be scrutinized by peers who have direct knowledge of the target market, the communicative motivation is qualitatively different from the motivation generated by instructor-only assessment. This principle can be extended beyond cross-national partnerships to include domestic industry practitioners, alumni networks, and community stakeholders as authentic audiences for student work.

From a curriculum design standpoint, the phased implementation structure adopted in this study (foundational input in Phase 1, collaborative problem-solving in Phase 2, culminating presentation in Phase 3) provides a replicable template for courses seeking to integrate active and collaborative pedagogies without sacrificing the content scaffolding that students need before tackling complex problems. The explicit sequencing of phases ensures that students arrive at the collaborative application tasks with sufficient disciplinary grounding to engage productively with them, addressing a common critique of PBL implementations in which insufficient content preparation leads to superficial rather than deep problem engagement.

5.4 Limitations and Future Directions

Several limitations qualify these findings. The sample (drawn from a single course at one Taiwanese institution and one U.S. community college) restricts the generalizability of results across educational systems, institutional types, and disciplinary contexts. The absence of a control group, a consequence of the naturalistic action research design, prevents strong causal attribution of the observed gains to any specific component of the intervention. Accordingly, the statistical paths reported in this study should be interpreted as evidence of theoretically grounded associations within an implemented pedagogical design rather than as definitive experimental proof of causal effects. Future research should address this through randomized or quasi-experimental designs that can isolate the independent contributions of CAR, PBL, and TBL.

Participant comparability also requires caution. The Taiwanese cohort was composed mainly of graduate business students, whereas the American cohort consisted primarily of undergraduate students from business and international studies programs. Although scaffolding, parallel tasks, common instruments, and rubric alignment were used to reduce this asymmetry, differences in academic level, prior disciplinary knowledge, and language-learning histories may have influenced both self-reported competencies and grade-based outcomes.

A related limitation concerns the confounding of pedagogical and contextual factors in the present design. The cross-national collaboration, the English-medium instruction, the specific disciplinary content of international marketing, and the three integrated pedagogies all co-varied, making it difficult to attribute outcomes to any single design feature with confidence. Systematic dismantling studies, in which versions of the course implement different subsets of the CAR–PBL–TBL integration, would help identify the active ingredients of the combined approach and their relative contributions to the four 4C dimensions. Such studies would also illuminate potential interaction effects, for example, whether TBL’s accountability structures are necessary for PBL to produce deep content learning, or whether PBL’s authentic problem contexts are necessary for TBL to generate the communicative engagement observed in this study.

Self-report measures introduce shared method variance and social desirability bias, particularly for constructs like cross-cultural understanding that have evaluative connotations. Behavioral and performance-based measures of the 4C dimensions (such as discourse analysis of team deliberations, scored intercultural scenario tasks, or longitudinal tracking of linguistic development) would strengthen construct validity and complement the self-report approach. Future studies might triangulate questionnaire data with instructor ratings of student performance on specific communicative tasks, linguistic analysis of student written and spoken output, and direct behavioral indicators of intercultural adaptability such as the quality and diversity of culturally informed arguments in the consultancy reports.

Finally, the cross-sectional design precludes examination of how competency development unfolds temporally. Longitudinal studies tracking cohorts across multiple semesters would illuminate how CAR–PBL–TBL environments shape longer-term trajectories of disciplinary expertise, communicative fluency, and intercultural identity. Such studies would be particularly valuable for testing the hypothesis that cognitive ability and cross-cultural understanding, the two non-significant mediators in the present study, develop along slower timescales than content knowledge and communication skills, requiring assessment intervals measured in academic years rather than semesters. They would also enable researchers to examine whether the competency gains observed in a single cross-national course persist, intensify, or attenuate as students move into subsequent courses and, eventually, into professional careers.

6. CONCLUSION

This study set out to examine how a systematically integrated pedagogical design combining Collaborative Action Research, Problem-Based Learning, and Team-Based Learning was associated with the four dimensions of Coyle’s CLIL 4C framework and, through those dimensions, with student learning performance in an English-medium (EMI/CLIL) business course. The evidence gathered from 81 students across two national cohorts indicates that the answer is substantially, though not uniformly, affirmative. Within this implemented design, students reported significant and practically meaningful enhancements across all four 4C dimensions, with particularly strong effects on cognitive ability and communication ability. The association of these gains with grades was, however, selective: content understanding and communication ability were significantly associated with learning performance, while cognitive ability and cross-cultural understanding were not. This selective mediation pattern invites a critical reconsideration of what English-medium business education is for and how its achievements should be measured.

The finding that grades fail to capture the full breadth of learning that the CAR–PBL–TBL design produces is not a failure of the pedagogy; it is a diagnostic signal about the adequacy of conventional assessment. Business schools aspiring to produce graduates who can operate effectively in international professional environments must invest not only in pedagogical innovation but in assessment innovation, developing instruments and rubrics that are sensitive to the cognitive depth, communicative sophistication, and intercultural sensitivity that high-quality bilingual business education actually cultivates. The present study provides an empirically grounded argument for undertaking that investment.

REFERENCES

- Banegas, Darío Luis, Anahí Pavese, Aurelia Velázquez, and Sandra María Vélez. 2013. Teacher professional development through collaborative action research: impact on foreign English-language teaching and learning. *Educational Action Research* 21(2):185-201. <https://doi.org/10.1080/09650792.2013.789717>
- Bartlett, Christopher A., and Sumantra Ghoshal. 1998. *Managing across borders: the transnational solution* (2nd ed.). Harvard Business School Press.
- Blasco-Arcas, Lorena, Isabel Buil, Blanca Hernández-Ortega, and Francisco Javier Sesé. 2013. Using clickers in class. The role of interactivity, active collaborative learning and engagement in learning performance. *Computers and Education* 62:102-110. <https://doi.org/10.1016/j.compedu.2012.10.019>
- Brown, Catherine A., and Hilda Borko. 1992. Becoming a mathematics teacher. In *Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (pp. 209-239). Macmillan Publishing Co, Inc.
- Carr, Wilfred, and Stephen Kemmis. 1986. *Becoming Critical: Education, Knowledge and Action Research*. Routledge. <https://doi.org/10.4324/9780203496626>
- Chang, Artemis, Julie Duck, and Prashant Bordia. 2006. Understanding the Multidimensionality of Group Development. *Small Group Research* 37(4):327-350. <https://doi.org/10.1177/1046496406290564>
- Chen, Guo-Ming, and William J. Starosta. 2000. The development and validation of the intercultural sensitivity scale. *Human Communication* 3:1-15.
- Chin, Wynne W. 2010. How to Write Up and Report PLS Analyses. In Vincenzo Esposito Vinzi, Wynne W. Chin, Jörg Henseler, and Huiwen Wang (Eds.), *Handbook of Partial Least Squares: Concepts, Methods and Applications* (pp. 655-690). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-32827-8_29
- Coyle, Do. 2005. *CLIL: Planning Tools for Teachers*. The University of Nottingham, School of Education. Nottingham. https://tuxdoc.com/download/clil-planning-tools-for-teachers-do-coyle-edited-version_pdf
- Coyle, Do, Philip Hood, and David Marsh. 2010. *CLIL: Content and Language Integrated Learning*. Cambridge University Press.
- Dalton-Puffer, Christiane. 2007. *Discourse in Content and Language Integrated Learning (CLIL) Classrooms*. John Benjamins Pub. <https://doi.org/10.1075/llt.20>
- Dalton-Puffer, Christiane. 2011. Content-and-Language Integrated Learning: From Practice to Principles? *Annual Review of Applied Linguistics* 31:182-204. <https://doi.org/10.1017/S0267190511000092>
- Dalton-Puffer, Christiane, Tarja Nikula, and Ute Smit. 2010. *Language Use and Language Learning in CLIL Classrooms*. John Benjamins Publishing Company. <https://books.google.com.tw/books?id=IT0zAAAAQBAJ>

- Deardorff, Darla K. 2006. Identification and Assessment of Intercultural Competence as a Student Outcome of Internationalization. *Journal of Studies in International Education* 10(3):241-266. <https://doi.org/10.1177/1028315306287002>
- Dolmans, Diana H. J. M., Sofie M. M. Loyens, Hélène Marcq, and David Gijbels. 2016. Deep and surface learning in problem-based learning: a review of the literature. *Advances in Health Sciences Education* 21(5):1087-1112. <https://doi.org/10.1007/s10459-015-9645-6>
- Ellis, Robert A., and Peter Goodyear. 2010. *Students' experiences of e-learning in higher education: the ecology of sustainable innovation*. Routledge. <https://doi.org/10.4324/9780203872970>
- Fennema, Elizabeth, and Megan Loef Franke. 1992. Teachers' knowledge and its impact. In *Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (pp. 147-164). Macmillan Publishing Co, Inc.
- Fornell, Claes, and David F. Larcker. 1981. Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research* 18(1):39-50. <https://doi.org/10.1177/002224378101800104>
- Haidet, Paul, Ruth E. Levine, Dean X. Parmelee, Sheila Crow, Frances Kennedy, P Adam Kelly, Linda Perkowski, Larry Michaelson, and Boyd F. Richards. 2012. Perspective: Guidelines for Reporting Team-Based Learning Activities in the Medical and Health Sciences Education Literature. *Academic Medicine* 87(3):292-299. <https://doi.org/10.1097/ACM.0b013e318244759e>
- Hair, Joseph F., Jr., Rolph E. Anderson, Ronald L. Tatham, and William C. Black. 1998. *Multivariate data analysis* (5th ed.). Pearson Education Inc.
- Hair, Joseph F., G. Tomas M. Hult, Christian M. Ringle, and Marko Sarstedt. 2017. *A primer on partial least squares structural equation modeling (PLS-SEM)* (2nd ed.). Sage.
- Hair, Joseph F., Christian M. Ringle, and Marko Sarstedt. 2011. PLS-SEM: Indeed a silver bullet. *Journal of Marketing Theory and Practice* 19(2):139-152.
- Hattie, John, and Helen Timperley. 2007. The Power of Feedback. *Review of Educational Research* 77(1):81-112. <https://doi.org/10.3102/003465430298487>
- Henseler, Jörg, Christian M. Ringle, and Rudolf R. Sinkovics. 2009. The use of partial least squares path modeling in international marketing. In *Advances in International Marketing* (Vol. 20, pp. 277-319). Emerald Publishing Limited. [https://doi.org/10.1108/S1474-7979\(2009\)0000020014](https://doi.org/10.1108/S1474-7979(2009)0000020014)
- Hmelo-Silver, Cindy E. 2004. Problem-Based Learning: What and How Do Students Learn? *Educational Psychology Review* 16(3):235-266. <https://doi.org/10.1023/B:EDPR.0000034022.16470.f3>
- Hmelo-Silver, Cindy E., Ravit Golan Duncan, and Clark A. Chinn. 2007. Scaffolding and Achievement in Problem-Based and Inquiry Learning: A Response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist* 42(2):99-107. <https://doi.org/10.1080/00461520701263368>

- Ho, Sophia Shi-Huei. 2017. The Impact of Taiwanese College Students' Learning Motivation from Self-Determination Perspective on Learning Outcomes: Moderating Roles of Multi-Traits. *Educational Sciences: Theory and Practice* 17(3):1015-1034. <https://doi.org/10.12738/estp.2017.3.0406>
- Hofstede, Geert. 2001. *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations* (2nd ed.). Sage Publications.
- Hulland, John. 1999. Use of partial least squares (PLS) in strategic management research: a review of four recent studies. *Strategic Management Journal* 20(2):195-204. [https://doi.org/10.1002/\(SICI\)1097-0266\(199902\)20:2<195::AID-SMJ13>3.0.CO;2-7](https://doi.org/10.1002/(SICI)1097-0266(199902)20:2<195::AID-SMJ13>3.0.CO;2-7)
- Kember, David. 2000. Misconceptions about the learning approaches, motivation and study practices of Asian students. *Higher Education* 40(1):99-121. <https://doi.org/10.1023/A:1004036826490>
- Knight, Jane, and Hans de Wit. 1997. *Internationalisation of higher education in Asia Pacific countries*. European Association for International Education.
- Mayer, Richard E., and Cheryl I. Johnson. 2010. Adding Instructional Features That Promote Learning in a Game-Like Environment. *Journal of Educational Computing Research* 42:241-265.
- McKeachie, Wilbert J., and Marilla Svinicki. 2006. *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers* (12th ed.). Houghton-Mifflin.
- Mehisto, Peeter, David Marsh, and María Jesús Frigols. 2008. *Uncovering CLIL: Content and Language Integrated Learning in Bilingual and Multilingual Education*. Macmillan Education. <https://books.google.com.tw/books?id=amvGGAAACAAJ>
- Mevarech, Zemira R., and Bracha Kramarski. 1997. IMPROVE: A multidimensional method for teaching mathematics in heterogeneous classrooms. *American Educational Research Journal* 34(2):365-394. <https://doi.org/10.2307/1163362>
- Michaelsen, Larry K., Arletta Bauman Knight, and L. Dee Fink. 2004. *Team-Based Learning: A Transformative Use of Small Groups in College Teaching*. Routledge.
- Michaelsen, Larry K., and Michael Sweet. 2008. The essential elements of team-based learning. *New Directions for Teaching and Learning* 2008(116):7-27. <https://doi.org/10.1002/tl.330>
- Michaelsen, Larry K., Michael Sweet, and Dean X. Parmelee. 2008. *Team-based learning: small group learning's next big step*. Jossey-Bass.
- Noffke, Susan E., and Kenneth M. Zeichner. 1987. Action research and teacher thinking. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.
- Nunnally, Jum C. 1978. *Psychometric theory* (2nd ed.). McGraw-Hill.
- Paas, Fred G. W. C. 1992. Training strategies for attaining transfer of problem-solving skill in statistics: A cognitive-load approach. *Journal of Educational Psychology* 84(4):429-434. <https://doi.org/10.1037/0022-0663.84.4.429>

- Parmelee, Dean X., Dan DeStephen, and Nicole J. Borges. 2009. Medical students' attitudes about team-based learning in a pre-clinical curriculum. *Medical Education Online* 14:1. <https://doi.org/10.3885/meo.2009.Res00280>
- Pintrich, Paul R., and Elisabeth V. De Groot. 1990. Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology* 82(1):33-40. <https://doi.org/10.1037/0022-0663.82.1.33>
- Polya, George. 1957. *How to solve it: A new aspect of mathematical method* (2nd ed.). Doubleday.
- Savery, John R. 2006. Overview of Problem-based Learning: Definitions and Distinctions. *Interdisciplinary Journal of Problem-Based Learning* 1(1):9-20. <https://doi.org/10.7771/1541-5015.1002>
- Savin-Baden, Maggi. 2014. Using Problem-based Learning: New Constellations for the 21st Century. *Journal on Excellence in College Teaching* 25:197-219.
- Savin-Baden, Maggi, and Claire Howell Major. 2004. *Foundations of Problem-Based Learning*. McGraw-Hill Companies, Incorporated. <https://books.google.com.tw/books?id=tU7oAAAAQBAJ>
- Schmidt, Henk G., Jerome I. Rotgans, and Elaine H. J. Yew. 2011. The process of problem-based learning: what works and why. *Medical Education* 45(8):792-806. <https://doi.org/10.1111/j.1365-2923.2011.04035.x>
- Schraw, Gregory, and Rayne Sperling Dennison. 1994. Assessing metacognitive awareness. *Contemporary Educational Psychology* 19(4):460-475. <https://doi.org/10.1006/ceps.1994.1033>
- Springer, Leonard, Mary Elizabeth Stanne, and Samuel S. Donovan. 1999. Effects of Small-Group Learning on Undergraduates in Science, Mathematics, Engineering, and Technology: A Meta-Analysis. *Review of Educational Research* 69(1):21-51. <https://doi.org/10.3102/00346543069001021>
- Taylor, David, and Barbara Mifflin. 2008. Problem-based learning: Where are we now? *Medical Teacher* 30(8):742-763. <https://doi.org/10.1080/01421590802217199>
- Timperley, Helen, Linda Kaser, and Judy Halbert. 2014. *A Framework for Transforming Learning in Schools: Innovation and the Spiral of Inquiry*. Centre for Strategic Education. <https://books.google.com.tw/books?id=2eOsoAEACAAJ>

Chih-Cheng Lin
 English Taught Program in International Business
 Shih Chien University
 Taipei, Taiwan
edcclin@gmail.com

APPENDIX.

The survey was administered to students from both institutions in Week 18. All items were rated on a seven-point scale (1 = strongly disagree, 7 = strongly agree). The English version is presented below; a Chinese translation was provided to Taiwanese students.

The following statements ask about your experiences in this course, which integrated Problem-Based Learning (PBL), Team-Based Learning (TBL), and Collaborative Action Research (CAR). Please indicate your level of agreement with each statement.

Section A: Integrated CAR–PBL–TBL Instructional Design (adapted from Mayer and Johnson 2010)

A1. The integration of PBL, TBL, and CAR in this course is an enjoyable experience.

A2. The integration of PBL, TBL, and CAR in this course is an effective learning tool.

A3. The integration of PBL, TBL, and CAR in this course has enhanced my problem-solving and decision-making skills.

A4. The integration of PBL, TBL, and CAR in this course has had a positive impact on my learning.

A5. The integration of PBL, TBL, and CAR in this course is a satisfying experience.

Section B: Content Understanding (adapted from Paas 1992)

B1. I can summarize what I have learned from this course.

B2. I understand the main ideas covered in this course.

B3. I can apply the knowledge I have gained from this course to solve real-world problems.

B4. I can explain the concepts I have learned in this course.

Section C: Cognitive Ability (adapted from Schraw and Dennison 1994; Pintrich and De Groot 1990)

C1. I believe my ability to understand and retain course material has improved.

C2. My skills for analyzing and synthesizing information have improved.

C3. I have become better at solving complex problems.

C4. My ability to focus on course tasks has improved.

C5. My capacity to retain and recall course content has increased.

C6. My ability to plan and manage my time for coursework has improved.

Section D: Communication Ability (adapted from Chang et al. 2006; Ellis and Goodyear 2010)

D1. I have become more comfortable expressing my ideas in group discussions.

D2. My ability to express my ideas clearly has improved.

D3. My skills in understanding and interpreting others' communication have improved.

D4. My ability to listen effectively in group discussions has improved.

Section E: Cross-Cultural Understanding (adapted from Chen and Starosta 2000; Deardorff 2006)

E1. I now appreciate cultural differences in behavior and values more than before.

E2. My understanding of the customs and traditions of different cultures has improved.

E3. I feel more comfortable interacting with people from different cultural backgrounds.

E4. My ability to adapt to different cultural contexts has improved.

整合協同行動、問題導向與團隊導向之跨國合作教學：
以國際行銷管理課程為例

林志丞
實踐大學

本研究探討在跨國內容與語言整合學習（CLIL）情境中，結合協同行動研究、問題導向學習與團隊導向學習，與國際行銷管理課程學生學習成效之間的關聯。研究以臺灣與美國學生共 81 人為樣本，運用 PLS-SEM 分析 4C 構面之中介效果。結果顯示，學生在內容理解、認知能力、溝通能力與跨文化理解四個構面上均呈現顯著提升；其中僅內容理解與溝通能力與學期成績顯著相關，認知能力與跨文化理解則與學期成績無顯著關聯。研究指出，傳統成績評量可能不足以完整呈現英語授課（EMI/CLIL）商管教育中的認知與跨文化能力發展。

關鍵字：問題導向學習、團隊導向學習、協同行動研究、內容與語言整合學習、雙語教育、商管教育

DEVELOPING NON-NATIVE EFL TEACHERS' EMPATHY THROUGH LEARNING A NEW FOREIGN LANGUAGE

Linh Thuy Nguyen
Hanoi University

ABSTRACT

Teacher empathy has gained wider recognition across educational contexts for its crucial role in fostering positive learning environments and improving teaching quality. However, despite the affective nature of language acquisition, the specific nuances of teacher empathy in ELT and its development remain underexplored. This study investigates how the immersive experience of learning a new foreign language impacts the empathetic understanding and teaching practices of language teachers by using qualitative research design. The participants were six non-native EFL teachers working in four universities in Vietnam. Focus group, diaries, semi-structured interviews, and thematic analysis were used to accomplish the objectives of the study. The results revealed that learning a new language, in general, helped reactivate and develop these teachers' empathy towards their students and propelled them to adopt changes in their teaching practices. A major conclusion from this study was that engaging in continuous foreign language learning can foster dynamic empathy in non-native EFL teachers, leading to more empathic and effective pedagogical practices.

Keywords: teacher empathy, social-emotional competence, emotional intelligence, non-native EFL teachers

1. INTRODUCTION

Affective factors have long been a branch of language teaching research, with the focus mainly on anxiety and motivation (Barcelos and Coelho 2016). However, under the influence of positive psychology, a broader spectrum of emotions in the language teaching and learning process have attracted the attention of researchers, one of which is none other than empathy. Empathy can often be understood as “wearing someone else’s shoes.” In other words, it is about how people try to “understand the minds and emotions of others”, thus developing their relations with one another (Mercer 2016:92).

In education in general and in language teaching in particular, teacher empathy has gained significant attention. Empathy is believed to help build positive relationships between teachers and learners, laying the foundation for enhanced learner engagement and academic performance in the classroom (Aldrup et al. 2022; Roorda et al. 2017; Zhang 2022). Moreover, empathy also has a positive effect on teachers themselves as they experience less burnout and deal better with problems during their teaching process (Wink et al. 2021). In the context of language teaching research, teacher empathy can contribute to students’ language learning as it helps build a positive classroom environment where teachers are believed to better understand their students’ experiences and make necessary adjustments to raise their teaching quality (Willis et al. 2025).

In the field of English language teaching, non-native English-speaking teachers (NNESTs) occupy a unique position as both educators and experienced language learners. Their own histories of learning English often provide them with insights into learners’ challenges, making them potentially more sensitive to students’ linguistic and emotional struggles (Lipovsky and Mahboob 2010; Mullock 2010). As a result, NNESTs have frequently been associated with strengths related to empathy, awareness of learner difficulties, and the ability to anticipate common errors.

However, current research into teacher empathy has mainly conceptualized it as a pre-existing trait and focused on its effects rather than the process of its development. In the case of NNESTs, while they possess rich linguistic backgrounds, their experiences of learning additional languages in adulthood, under the conditions of professional

responsibility, time pressure, and identity shifts, remain largely unexplored in relation to empathy development. Understanding how such experiences shape teachers' empathy can provide new insights into the processes through which empathy is formed, transformed, and enacted in language teaching. Therefore, this study is set to investigate the formation of NNESTs' empathy. It is expected that this study could be a suggestion for teachers to develop their own empathy, thereby raising their teaching quality in the classroom.

2. LITERATURE REVIEW

2.1 Non-Native English-Speaking Teachers

A large number of English speakers today use it as a medium of international and intercultural communication in many non-native English contexts (Sharifan 2009), which means that they can be identified as non-native English speakers (NNEs) (Bolton 2006). It is estimated that around 80% of global English teachers are NNEs, and they can be called non-native English-speaking teachers (NNESTs) (Oxford and Jain 2010).

NNESTs have a number of advantages in delivering language teaching practices. Firstly, they can act as role models for learning English (Colmenero and Lasagabaster 2020) as their hands-on experience of learning English can provide them a deeper understanding of their students' potential struggle. This could be further enhanced if NNESTs and students share the same L1 and culture. They can better communicate and build closer rapport with students (Ma 2012). As for imparting mechanic knowledge, they can explain grammar and difficult points more easily to students (Walkinshaw and Oanh 2014) as they have a good command of metacognitive language awareness (Arva and Medgyes 2000). Lastly, they are familiar with the demands of the local educational system (Brown and Lee 2015), allowing them to better prepare their students to reach their goals.

In terms of empathy, NNESTs are believed to have greater empathy towards learners (Lipovsky and Mahboob 2010; Medgyes 1994). This empathy is rooted in their shared language learning journey (Mullock

2010). Students perceived their NNESTs as having empathy with them, since they had experienced the same difficulties as them in learning English. This perception stems from the belief that NNESTs know that learners “cannot understand language which we are learning easy” and that they “also know students’ feeling” (Lipovsky and Mahboob 2010).

2.2 Teacher Empathy

Howe defines empathy as an emotion with four major components: “as an affective reaction to the emotions of another; the cognitive act of adopting another’s perspective; a cognitively based understanding of other people; and the communication of such an understanding” (Howe 2013:14). This means that an empathetic individual is a person who is able to understand the minds and emotions of another, see things from that person’s point of view and tell that person they understand what the other person is feeling. As a result, empathy is one of the most important and beautiful emotions that humans have, leading to a growing interest in the role of empathy in the relationships between humans.

Under the influence of this trend, empathy has also garnered attention in the field of language teaching, leading a number of scholars to attempt to give their own definitions of teacher empathy. Tettegah and Anderson (2007) constructs teacher empathy as the ability to understand students’ perspectives, concerns and emotions. Meyers et al. (2019), based on the works of Batson (2009) and Segal (2011) claim that teacher empathy exists on a continuum, and it varies depending on the contexts that teachers are in. As for English language teaching, Willis et al. (2025) argue that empathetic teachers should be able to not only recognize and understand their students’ emotions but also respond to their both positive and negative experiences in learning a new language.

In positive psychology, empathy consists of three main types: emotional empathy, cognitive empathy and compassion or sympathy (Mercer 2016). Emotional empathy involves sharing or resonating with others’ emotions, especially when individuals respond to others’ distress based on similar experiences (Hodges and Myers 2007). Cognitive empathy refers to the ability to draw on prior knowledge or available cues to understand and infer another person’s mental and emotional state,

particularly in situations of suffering (Decety and Jackson 2004; Smith 2006). Compassionate empathy goes beyond understanding and feeling, encompassing a willingness to take intentional steps to reduce or ease others' suffering (Maxwell 2017). It can stem from cognitive empathy, emotional empathy (Ekman 2003), or a combination of both (Maxwell 2017). Below is a diagram showing the relationship between these types of empathy:

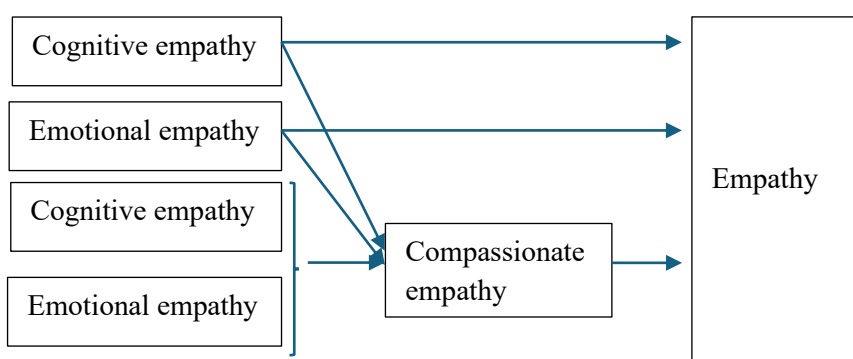


Figure 1. Different types of empathy in psychology.

This trichotomy has laid the foundation for the research of teacher empathy. However, many studies on teacher empathy only focus on the two first components of empathy i.e., cognitive (understanding students' emotions intellectually) and emotional empathy (sharing or resonating with them emotionally) (Aldrup et al. 2022; Zhang 2022). In the context of English language teaching, Willis et al. (2025) further expanded empathy to include reflective and enacted forms, with reflective empathy involving understanding learners through self-reflection on one's own language-learning experiences, while enacted empathy being the outward pedagogical expression of that understanding. The concept of reflective empathy builds on emotional empathy, but it is different from emotional empathy in the sense that the former encompasses both positive and negative experiences while the latter only includes negative ones. As for enacted empathy, it includes taking empathy-induced action, which makes it similar to compassionate empathy to a certain extent. However, it differs from compassionate empathy, with enacted empathy arising directly from

reflective empathy while compassionate empathy coming from either cognitive empathy, emotional empathy or a combination of both (Willis et al. 2025). The relationships between these types of empathy are shown below:

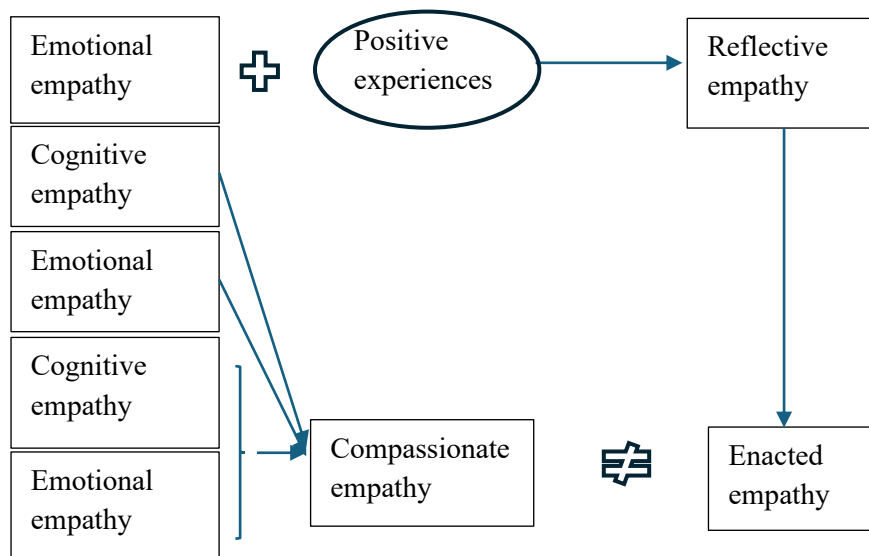


Figure 2. The relationship between different types of empathy in education.

Teacher empathy is influenced by a variety of variables. Gender and attitudes towards students play a role in shaping teacher empathy (Parchomiuk 2019; Pidbutska et al. 2021). Another factor is personality traits like teacher agreeableness, which also contributes to the formation of teacher empathy (Klassen et al. 2017). Moreover, teachers' professional identity has a positive impact on teacher empathy (Zhu et al. 2019). Emotional intelligence is another variable that has been confirmed by Salem and Tabatabaei (2015). Therefore, it is safe to say that teacher empathy is a complex psychological construct under the influence of many factors.

2.3 Teacher Empathy in Education

In education, teacher empathy helps build stronger relationships between teachers and students. Aldrup et al. (2022) and Roorda et al. (2017) found that empathy showed the most consistent positive evidence in the domain of emotional support, particularly in contexts where teachers respond to students' emotional needs, but the strength of this effect varies across contexts and research designs as not all studies found consistent effects (Aldrup et al. 2022). However, as for other aspects of teacher-student interaction, teacher empathy seems to have little significant impacts. Studies have shown that the relationships between teacher empathy and classroom management and teacher empathy and instructional support are generally weak (Aldrup et al. 2022).

With regard to the effects of teacher empathy on students' outcomes, research findings have been inconsistent. For students' cognitive outcomes, empathy has mixed direct influence, with Curci et al. (2014) concluding that empathy generally did not predict students' academic performance while Franklin (2014) and Latchaw (2016) finding that empathic teachers were associated with better reading. Turning to psychological student outcomes, most studies examining this aspect reported no significant associations, making this another domain with few positive impacts (Aldrup et al. 2022).

Besides the benefits for students, teacher empathy has been found to exert positive effects on educators themselves. Cooper (2002) and Goroshit and Hen (2016) confirm that teachers who have a high level of empathy are more effective in delivering pedagogical practices and have greater self-efficacy, allowing them to better deal with any challenges in their classrooms. Furthermore, teacher empathy also helps reduce work-related burnout and foster problem-solving strategies for teachers (Wink et al. 2021).

2.4 Teacher Empathy in Language Teaching and Learning

Research into the role of empathy dates back to the 1970s when Guiora et al. (1972) found that students with more empathy could better mimic a native speaker's pronunciation. Since then, several other studies have also investigated the relationship between empathy and L2 learners'

pronunciation aptitude (Dewaele and Wei 2012; Ozanska-Ponikwia and Dewaele 2012; Rota and Reiterer 2009).

Teacher empathy has also been concluded to have positive correlations with students' language learning. Empathy, kindness, and tolerance are vital for building positive relationships and classroom climates that enhance motivation and achievement (Gkonou and Mercer 2017; Ocampo Jaramillo et al. 2025). In addition, empathic teachers create emotionally safe learning environments, reduce anxiety, and boost learners' self-confidence and engagement (Sadoughi and Hejazi 2021). Another benefit is teacher empathy and rapport predict higher student engagement in EFL settings (Derakhshan 2021; Hashim et al. 2014; Teo et al. 2022). Lastly, empathic teachers have a deeper understanding of their students' learning experiences, thus attuning to their learning needs by recalling their experiences as a language learner and using those experiences to transform their teaching practices (Willis et al. 2025).

In the context of language teaching and learning in Viet Nam, there has been a limited body of research into empathy. Hieu and Thao (2024) found that Vietnamese teachers showed a dynamic shift in their perception and acceptance of empathy in teaching, and they found it challenging to integrate empathy in the curriculum. Le et al. (2024) explored different strategies Vietnamese teachers used to showcase their empathy towards students, including providing customized feedback, listening more empathetically, creating empathetic group dynamics, and integrating student narratives into their exchanges with students.

The existing Vietnamese studies highlight how teachers perceive and enact empathy in their teaching practices (Hieu and Thao 2024; Le et al. 2024), but they do not address how teachers actually develop empathy. Their focus remains on empathetic teaching behaviors rather than the experiential or developmental sources from which empathy emerges. Meanwhile, the literature on NNESTs suggest that they may develop empathy in unique ways, yet the literature has not examined how NNESTs cultivate empathy over time. Research outside Vietnam indicates that teachers often use their past language-learning experiences to empathize with students (Ellis 2016; Moodie 2016; Willis et al. 2025), but no studies have explored whether current or new foreign language learning can actively foster empathy. This gap is particularly salient in the Vietnamese

context, where empathy is valued but its developmental pathways remain unexamined. These unresolved issues raise the need to investigate how learning a new foreign language may contribute to NNESTs' empathy development, leading to the following research questions:

Q1. How does learning a new foreign language contribute to NNEST' empathy towards students?

Q2. How does teacher empathy towards students contribute to EFL teaching practices?

3. METHODOLOGY

A qualitative research design was chosen because of its appropriateness to the nature of the research topic, which is to shed light on how learning a new foreign language can develop NNESTs' teaching empathy. This topic requires the researchers to gain access to the participants' inner thoughts and beliefs, rendering the employment of a qualitative approach most suitable because qualitative research is often used to find out personal feelings, experiences and viewpoints (Dornyei 2007).

3.1 Setting and Participants

The participants were six Vietnamese EFL teachers working in four different universities in Viet Nam, who were purposefully recruited for this study. Each teacher had at least three years of experience in teaching EFL and had to achieve at least B2 level as an enrollment requirement for their PhD programs in Viet Nam. Before the study, each teacher had a different level of L3. One was [a](#) beginner, two had A2 level, and the other three had B1 level. The reason for the difference in their beginning levels was that some had to take at least 3 semesters of mandatory second foreign language learning during undergraduate study to earn their bachelor's degrees in English Linguistics or English Language Teaching, but most of them did not use the L3 after they graduated. Below is a table that summarized the profile of the participants.

Table 1. Participants' profiles.

Teacher	Years of teaching experience	First language	Second language	Third language and starting level
T1	4 years	Vietnamese	English	Chinese - beginner
T2	3 years	Vietnamese	English	Chinese – A2
T3	5 years	Vietnamese	English	Korean – B1
T4	3 years	Vietnamese	English	Korean – A2
T5	8 years	Vietnamese	English	Korean – B1
T6	5 years	Vietnamese	English	Japanese – B1

3.2 Data Collection Instruments

To explore the teachers' empathy development, focus group discussions, diaries and semi-structured interviews were employed to gather data and there are several reasons for the use of these methods in the process of data collection. Firstly, focus group discussions allowed the researcher to gather a relatively large amount of qualitative data in an economical way while allowing them to think together, inspire one another, and react to the emerging issues (Dornyei 2007), which was particularly helpful in finding out the teacher participants' inherent empathy towards their students because the teachers all shared similar English learning and teaching journeys. Secondly, written records like diaries could act as a time-effective means to obtain research participants' language because researchers do not need to transcribe before conducting data analysis (Creswell and Creswell 2018). Another reason is data from the diary entries can pave the way for later semi-structured interviews. After the teachers' biweekly entries were collected, data could be obtained immediately, thereby offering the researchers insights into their emotional experiences. The insights could lay the foundation for the subsequent interviews as gaps in the participants' language had been identified. The last reason for using this research design is using focus group discussions, diaries and semi-structured interviews could raise research validity. What the participants meant in their entries could be different from what the

researchers interpreted, making the subsequent interviews able to help the two parties understand each other. Therefore, the use of these three research instruments can boost the internal validity of the research (Johnson and Christensen 2016).

Before the research study was carried out, the participants signed a consent form to confirm their agreement to participate in the study.

3.3 Procedure

The data was collected in the first quarter of 2025. The teachers were first invited to join the study where they would share their emotional experiences in learning a second foreign language after they had become language teachers. They were also informed that the study would last 12 weeks, with their FGD, diaries and semi-structured interviews being conducted in English, and all would be confidential. They also knew that all the collected data would not be stored after the study was finished, and they signed consent forms to formally agree to participate in the study.

Firstly, before the teachers started their L3 courses, a two-hour FGD with all six of them took place to introduce the study, outline their roles and tasks and walk them through the procedures of data collection. In this FGD, they also discussed their struggles in learning the first foreign language i.e., English, their experiences in learning the second foreign language when they were university students, and the empathy they had towards their students. The participants were asked to use the diary guide to report their emotional experiences in their learning journey.

In the diary-writing process, the teachers were asked to record the lesson content, the opportunities and challenges in completing the assignments, and their emotions during this journey. Their diary entries were emailed every three weeks to the researcher. The total number of diary entries that the researcher received was 24 entries, all of which were then analyzed to identify themes, which were later used to develop interview guides for semi-structured interviews.

The interviews were held at the end of their courses, where the teachers were invited to elaborate on any emotional experiences they had during the whole course. The interviews were video-recorded for later transcribing process. Each interview lasted between around 45 to 60

minutes.

3.4 Data Analysis

For the qualitative analysis, reflective thematic analysis was employed to examine the data because it is suitable to understand personal experiences, thoughts or behaviors in a data set (Braun and Clarke 2022). The analysis mainly adopted an inductive approach, where themes were developed from the data itself rather than being based on pre-existing theories (Braun and Clarke 2022). This allowed the participants' experiences as both language learners and teachers to shape the development of the analysis. At the same time, the analysis adopted a latent orientation, going beyond the surface meanings of participants' accounts to interpret underlying emotions, identity shifts, and cognitive processes. For example, expressions of difficulty or discomfort were examined in relation to deeper constructs such as vulnerability, learner identity, and empathy development. The study was situated within an experiential qualitative framework, aiming to capture and understand participants' lived experiences as they navigated dual roles as teachers and learners. While the primary focus was on participants' perspectives, the analysis also incorporated elements of critical reflection, particularly when examining issues related to power, classroom practices, and emotional impact. Finally, the analysis was carried out within a realist epistemological stance, which assumes that participants' accounts represent their actual experiences and perspectives (Braun and Clarke 2022). This study sought to represent the meanings participants attributed to their experiences while acknowledging that these meanings are shaped by their learning contexts.

The data analysis process first started with the coding and theming of the FGD for later development of the interview guide. Through data familiarization, the researcher identified many analytically interesting points, which yielded hundreds of codes. Many of these codes were only slightly different from one another, so they were reduced into a smaller number. The codes were then grouped into three major themes: L2-learning-induced shame, L3-learning motivation and pre-existing empathy. For this paper, the researcher focused on developing the analysis

of the data connected to ‘empathy’ themes.

The process then continued in 4 three-week cycles, each of which commenced with the coding and theming of triweekly diary entries. Every entry underwent the same coding procedure as the FGD, resulting in hundreds of codes, which laid the foundation for the major theme: newly emerging empathy. The themes, combined with those from the FGD, were later used as a basis for the semi-structured interview guide. The interviews were then conducted, video-recorded and transcribed. From the obtained transcripts, the researcher identified themes from these discussions, which were later combined and connected with the previous themes from the diary entries to create overarching themes. This time, there were three main themes: pre-existing empathy, newly emerging empathy, and empathy in action.

For each coding and theming time during this study, the six-step process of reflexive thematic analysis by Braun and Clarke (2022) was adopted. The specific six steps are outlined in Table 2.

Table 2. Coding and theming process (Braun and Clarke 2022).

Step 1	Familiarizing with data: The researcher read and reread the whole dataset at least twice to get used to the collected data.
Step 2	Coding: The researcher identified potentially interesting, relevant or meaning segments of data for the research questions and gave them code labels. Table 3 illustrates this step.
Step 3	Generating initial themes: The researcher compiled clusters of codes sharing a core idea, the products of which were candidate themes (Table 4).
Step 4	Developing and reviewing themes: The researcher assessed the fit of candidate themes to the data and their suitability to the research questions. At this stage, the researcher combined overlapping themes, split bigger themes into smaller themes, and discarded irrelevant ones.
Step 5	Refining, defining and naming themes: The researcher made sure that the themes were relevant to the questions and built around a core concept, and the researcher finalized the name of each theme.

Step 6	Writing up: The researcher finessed and finished the analysis into two different products at different stages of this study i.e., the interview guide and the analysis report. After developing themes from the collected data from the FGD and the diary entries, the researcher produced an interview guide for the subsequent interviews. Once the data from the interviews had been coded and themed, the researcher engaged in writing the final analysis report.
--------	--

Table 3. Sample of collected data and code labels.

Data	Codes
T4: “You know the most hurtful thing is when your teacher calls you someone with zero ability to learn foreign languages just because you keep making the same errors. Well, I just forgot to conjugate a verb in my entire diagram report, and the error was at a pretty basic level like A1 or A2 while I was aiming for B2. And my teacher, he just suddenly crashed out and labeled me with that ugly thing. Never in my life I felt that angry and embarrassed. What even worse is that I have C2 level in English. How can he give me such an ugly comment? That was when I realize the importance of emotional regulation and perspective taking in teaching to show some empathy towards your own students.”	Teacher as evaluator of learner’s worth Harmful teacher judgments Identity attack through feedback Single error magnified Basic mistakes as unacceptable High expectations placed on advanced learners Perfectionism in language learning Discrepancy between goals and performance Teacher emotional dysregulation Negative emotions spilling into teaching Teacher behavior as unpredictable Shame in language learning Emotional injury Loss of face Coexistence of anger and embarrassment Language learning vulnerability Threatened self-esteem Public embarrassment Social humiliation Fear of negative evaluation

	Classroom as an unsafe space Learning as a socially risky activity Competent learner identity disrupted Expertise does not protect against vulnerability High achievers can struggle too Previous success does not guarantee future competence Moral judgment of teacher behavior Perceived unfairness Expectation of respectful treatment Learners deserve dignity Transformative learning experience Empathy triggered by suffering Learning through negative experience Teacher emotions affect learners Emotional self-control as professionalism Emotional responsibility of teachers Preventing emotional harm Awareness of learner emotions Empathy as understanding Empathy as professional responsibility Suffering transformed into care Compassion born from shared experience Breaking the cycle of harmful teaching Desire not to reproduce past
--	---

	trauma
--	--------

Table 4. Sample of candidate themes.

Codes	Candidate themes
Emotional injury Shame in language learning Loss of face Coexistence of anger and embarrassment Threatened self-esteem	Emotional suffering
Public embarrassment Social humiliation Fear of negative evaluation Classroom as an unsafe space Learning as a socially risky activity	Public exposure and social threat
Teacher emotional dysregulation Negative emotions spilling into teaching Identity attack through feedback Teacher as evaluator of learner's worth Harmful teacher judgments Teacher behavior as unpredictable	Harmful teacher behavior

4. RESULTS AND DISCUSSION

The researcher constructed three major themes, which were evident across the accounts of the participants. The first theme, *Pre-existing empathy*, shows the empathy that the teacher participants had before starting their L3 learning journey. The second theme, *Newly emerging empathy*, is the empathy activated through L3 learning, and the last one, *Empathy in action*, shows how the participants transform their empathy into pedagogical changes.

4.1 Pre-Existing Empathy

Before joining the study, the teachers had had a certain degree of empathy towards their students, which was the result of their L2 learning

journey i.e., English. This type of empathy was reflected in two main factors: shared L1 as an empathic bridge and unfavorable instruction.

4.1.1 Shared L1 as an empathic bridge

All of the participants teach Vietnamese students, who share their mother tongue, so they understood the actual challenges for Vietnamese learners to succeed in English language learning, especially in terms of grammar and pronunciation, as T1 shared:

“There are certain differences in the word order or writing style between Vietnamese and English, so of course, as a former English language learner, I noticed the differences and try to remind my students of them. Not all of the students can remember what I’ve told them, and I have to correct the same errors again and again, but sometimes I think I used to be just like them so yeah ... I get how they feel.”

“Well, certain sounds only exist in English, but not in Northern Vietnamese, like th- sound or sh- sound and even some vowels might sound similar at first, but actually they have a tiny bit of difference. I know the difficulties in pronunciation like this, so I understand my students better when teaching them speaking.” (T5)

T1 and T5 attributed their empathy for their students to their shared native language by specifically discussing the challenges that their students might encounter, such as syntactic rules, style, and pronunciation. These responses support what has been found in previous research, where NNESTs are perceived to be more empathetic with their students as a result of their previous learning experiences (Lipovsky and Mahboob 2010; Medgyes 1994). Moreover, they are also in line with the term “reflective empathy” introduced by Willis et al. (2025) where the first-hand learning experiences of an additional language plays a role in forming language teachers’ empathy.

4.1.2 Unfavorable instruction of previous language teachers

Another aspect of the teachers’ pre-existing empathy stemmed from unfavorable instruction of their previous language teachers. Some teaching practices were so intimidating that some of the participants still

remembered the situations vividly, which partly fueled the empathy for their students.

“One time, my English teacher corrected my pronunciation on the spot while I was speaking in front of the class. I felt really embarrassed at that time because my friends were looking at me, some even laughed at me. Another time was when she [the former English teacher] preferred us to pronounce in the American way. You know the ‘a’ sound as in ‘apple’, she wants us to pronounce like the Americans, but at that time, I just said it in the British way, and she called me wrong. I was very very angry and embarrassed. I know it’s not wrong, it’s just that the way I speak was different from what she wanted. So I always try to be super careful when correcting my students’ speaking errors because I don’t want them to feel what I felt.” (T4)

T4’s account reveals how emotionally painful experiences with former English teachers contributed to the development of empathy. Her recollection suggests that language learning was not merely cognitively demanding but also socially and emotionally risky. Being corrected publicly and laughed at by classmates transformed the classroom into an unsafe space where mistakes became sources of humiliation and loss of face. This is in line with current literature on immediate corrective feedback in public, which can leave a negative emotional experience on the feedback receivers i.e., students because it is potentially anxiety-inducing and intimidating to learners (Ellis 2021; Oliver and Adams 2021), requiring teachers to be far more careful when using it. More importantly, her narrative illustrates how personal suffering became a source of empathy. Rather than reproducing the practices that had caused her emotional harm, T4 became increasingly aware of the emotional consequences of feedback and more empathetic when their students made similar mistakes (Willis et al. 2025).

During the L3 classes, the pre-existing empathy was further reactivated and reinforced because many experiences were similar to what they first experienced when learning English. As T2 shared:

“When I read some Chinese characters out loud, I also got corrected by my teachers right away. I felt a bit horrible at that time because it reminds me of how I [her English pronunciation] was corrected when I was in high school. This just makes me more sensitive when it comes to

correcting someone else's pronunciation." (T2)

The immediate correction she received in her Chinese class triggered an emotional memory of her own pronunciation struggles as an English learner in high school, revealing how language-learning contexts can evoke previous affective responses. This process aligns partially with Willis et al. (2025), who describe teachers drawing on their past learning experiences to cultivate reflective empathy, but T2's experience goes beyond retrospection by showing how current L3 learning reopens and reinforces earlier vulnerabilities. This reactivation mechanism appears particularly salient for NNESTs, whose multilingual learning histories position them to resonate with learners' anxieties (McNeill 2005). Moreover, this also helps to answer the first research question as learning a new foreign language can reactivate and reinforce empathic understandings rooted in teachers' previous language-learning experiences.

4.2 Newly Emerging Empathy

Besides pre-existing empathy, newly emerging empathy was another theme that can be seen in the interviews with the participants. Newly emerging empathy can be defined as a new type of empathy that only appears when the participants learned a new foreign language as both a student and an in-service teacher. Under newly emerging empathy, several subthemes surfaced:

4.2.1 Age of learning

All six teachers learned English when they were in primary school, and they had formal lessons up until they were sophomores at university, which was different from the age at which some of their students took their language courses. Therefore, taking a new language course when they are adults gave them a chance to develop their empathy towards their students, as in the case of T3:

"I mostly teach adult learners who decide to learn English properly after they have graduated from their university, so to be honest, I did not know what it felt like to study a new language when you're an adult. But this course got me to realize how hard it is for adult learners to learn a new

language. It's really different from when you are at school, when you don't have to do anything except study. You don't have to worry about money or family or children. You don't have that kind of struggle and all that pressure on you. I just suddenly realized how hard it must be for my students to come to my class every week."

T3's narrative reveals how teachers' own educational histories can unintentionally create blind spots in their understanding of learners. Her experience as an L3 learner exposed the unequal conditions under which language learning takes place and challenged the assumption that all learners have similar capacities to devote themselves to study. Students' attendance and persistence were no longer interpreted as ordinary expectations but as evidence of effort and sacrifice. Empathy, therefore, emerged not simply from understanding learners intellectually, but from experiencing firsthand the tensions between learning and adult life, transforming previous assumptions about what students must endure to continue their language learning journeys. This finding supports the view that empathy is dynamic and shifting (Willis et al. 2025), yet it extends existing views by demonstrating that empathy can emerge through new learning situations rather than solely through recollection of past struggles.

T6 added to this:

"... the fear of making mistakes when you're a fully functional adult just hits different. Suddenly, you feel so dumb in a class, especially when you're the oldest [in the class]. Maybe some of my older students are also feeling this way."

In the case of T6, age functions as an important factor in the development of newly emerging empathy. As she describes, the fear of making mistakes "as a fully functional adult" generated a sense of embarrassment, particularly because she was the oldest learner in the room. Her experience supports Aldrup et al.'s (2022) view of empathy as contextually sensitive, yet it extends existing work by demonstrating that age-related vulnerability can trigger empathic insight. While Linville (2015), Moodie (2016), and Willis et al. (2025) emphasize retrospective empathy drawn from past learning histories, T6's account shows empathy can emerge in real time. Thus, her narrative reveals age-related vulnerability as a unique mechanism through which new foreign language learning can cultivate fresh understandings of learner experience.

4.2.2 Shared linguistic understanding with students having different L1

Learning a new foreign language enables the participants to enhance their connection with their foreign students, who do not share their L1. T4 commented:

“Learning Korean gave me a chance to understand my Korean students more deeply. I learned the basics of Korean when I was in my first year of university, but it has been 10 years since then, so I don’t remember much. But when I retook Korean courses, I understood better why my Korean students make certain types of syntactic errors.”

T1 added:

“This is the first time I’ve studied Chinese, so it’s great to understand the challenges that my Chinese students have in learning English. Like you can compare and contrast the languages and see what’s preventing them from reaching their targets.”

T4’s and T1’s reflections reveal how learning a new foreign language can cultivate a form of empathy grounded in understanding the structural relationship between languages. By revisiting Korean and learning Chinese for the first time, both teachers gained deeper insights into why their Korean and Chinese students produced certain syntactic patterns in English. This shows a cognitive dimension of empathy: understanding learners’ errors not as deficiencies but as natural products of L1 transfer, making it align with cognitive empathy (Decety and Jackson 2004; Smith 2006). This finding shows that empathy encompasses cognitive perspective-taking, yet it goes further by showing how linguistic perspective-taking emerges through L3 learning. Furthermore, T4 and T1’s accounts demonstrate that empathy can develop through active comparison between language systems, confirming the findings of Nemtchinova (2010) stating that NNESTs’ multilingual experience can arouse empathy and understanding towards their students. In this way, L3 learning becomes a pathway not only to emotional sharing but also to the understanding of learner difficulty.

4.2.3 Experiential knowledge of instructed L3 learning

Taking on the role of both a teacher outside the L3 class and a learner in the L3 class at the same time drove the participants to empathize with their students in several different ways. Firstly, three participants (T3, T5, T6) noted how being in an intensive course helped them understand their students.

“Choosing an intensive course to quickly achieve my target makes me realize how stressful intensive courses can be for my students. I’ve never had to take any intensive courses to get any kind of language certificates before, so it’s really new and overwhelming at the same time. Combining this with the work pressure ... I think it’s really hard to learn a new language when you have adult responsibilities.” (T3)

T3’s narrative challenges the assumption that language learning is simply a matter of effort or motivation. Through experiencing the pressures of an intensive course while juggling professional responsibilities, she became aware that learning takes place under unequal conditions and that adult learners often carry burdens outside the classroom. In this sense, empathy emerged through shared experiences, leading T3 to reinterpret students’ struggles not as signs of a lack of commitment but as the consequences of balancing multiple responsibilities. T3’s developing empathy illustrates how experiential learning environments can heighten teachers’ awareness of learners’ challenges (Willis et al. 2025).

“I used to be the top student in English back when I was in school, but it feels different when you’re no longer the top student in class. The peer pressure can be pretty tough (for students), especially when sometimes you’re in the bottom set of the class.” (T5)

T5’s account highlights how newly emerging empathy can develop through a shift in academic status, allowing her to gain insight into the peer pressure her own students may face, especially those consistently placed in lower proficiency groups. Her narrative shows that empathy can arise through the emotional experience of losing competence and navigating social comparison.

“...cause that I’m also a teacher, I notice the way my Japanese teacher teaches in class. Some of his practices are similar to mine, and among those, some works, some don’t. I just think that may be this is how my students are feeling all the time, that they need more explanation or

repetition because it's quite hard to take in everything in just a few times [of repetition] ...” (T6)

Seeing herself reflected in another teacher enabled T6 to realize that her teaching practices might not adequately support her learners. This role reversal allowed her to understand how limited explanation or insufficient repetition can overwhelm learners, prompting her to consider that her own students may often face similar difficulties. This realization can be seen as a direct example of ‘cognitive empathy’, which allows teachers to understand students’ perspectives and intellectually infers their mental states (“may be this is how my students are feeling all the time”) (Decety and Jackson 2004; Smith 2006).

4.2.4 Negative affective incidents

T4 shared her experience in her Korean class:

“You know the most hurtful thing is when your teacher calls you someone with zero ability to learn foreign languages just because you keep making the same errors. Well, I just forgot to conjugate a verb in my entire diagram report, and the error was at a pretty basic level like A1 or A2 while I was aiming for B2. And my teacher, he just suddenly crashed out and labeled me with that ugly thing. Never in my life I felt that angry and embarrassed. What even worse is that I have C2 level in English. How can he give me such an ugly comment? That was when I realize the importance of emotional regulation and perspective taking in teaching to show some empathy towards your own students.”

The deeper story underlying T4's narrative is that language learning errors are not merely linguistic events but experiences that can threaten learners' self-confidence and self-esteem under teachers' judgments and emotional outbursts, echoing findings by Roothoof and Breeze (2016). Despite being a highly competent language user as she has C2 level in English, T4's experience demonstrated that expertise does not protect learners from shame and vulnerability. This incident also exemplifies the importance of teacher emotional regulation noted by Aldrup et al. (2023), which T4 explicitly identifies as lacking in her instructor's response.

Furthermore, the incident transformed her understanding of what it means to be a teacher. Rather than viewing teaching primarily as the

transmission of knowledge, she came to see it as an ethical and emotional practice requiring self-regulation and perspective-taking. In this sense, empathy emerged through suffering, with personal wounds becoming a catalyst for care and a commitment to preventing students from experiencing similar emotional harm. This suggests that developing empathy is a dynamic and ongoing process happening in real time, developing what Willis et al. (2025) found about ‘reflective empathy’, which stems from relying on past language-learning experiences.

These abovementioned findings help answer the first research question on whether learning a new language contributes to NNESTs’ empathy towards students. Learning a new foreign language creates opportunities for teachers to experience their students’ realities firsthand, leading to new forms of empathy that arise through role reversal, shared struggles, and pedagogical recognition.

4.3 Empathy in Action

As the teachers worked and studied simultaneously, some expressed that once they could take their students’ perspectives, they had tried to adopt changes in their daily teaching practices, which can be termed empathy in action. This type of empathy is, to some extent, similar to ‘enacted empathy’ by Willis et al. (2025) as they both stem from the desire to carry out changes to teaching practices based on teachers’ past language learning experiences, yet they differ in their temporal orientation. Willis et al.’s (2025) participants drew on earlier language-learning experiences, the teachers in the present study enacted empathy while they were still undergoing the experiences that shaped their empathic understanding, resulting in a much shorter temporal distance between experience and action.

This empathy was shown in the following aspects:

4.3.1 Giving feedback with care

“I think there are different ways to correct a students’ pronunciation. In my case, I think my teacher could’ve done differently like feedback to me in private or at least wait until I finish my talk” (T2)

This extract shows a change in how correction is understood. Rather than viewing feedback as an act aimed at maximizing accuracy, T2 came to understand it as a practice with emotional consequences for learners. Her suggestions regarding private feedback and delayed correction indicate a growing awareness that mistakes should not expose learners to unnecessary embarrassment or disrupt their attempts to communicate. Consequently, protecting learners' image became more important than correcting errors as quickly or efficiently as possible. This means that empathy reshaped her understanding of feedback.

4.3.2 Modifying teaching practices

"I have put in more guided and focused practice for my lower-level students. I used to assume that they can easily rewrite an essay with the sample answer I provide them, but only when I was wearing their shoes, I realized how hard it is to mimic a sample essay where the writer can flexibly use their language. The repertoire of lower-level students' not that big and it can be stressful for them." (T4)

T4's narrative demonstrates how empathy acquired through L3 learning can directly inform changes in instructional design. Her own experience as a language learner exposed the gap between understanding what good performance looks like and having the linguistic resources to achieve it. As a result, she came to recognize that learners' struggles often stem from limited resources rather than a lack of effort or motivation. This shows how empathy becomes an agent of pedagogical changes, prompting teachers to adjust expectations, refine task design, and provide more flexible and realistic pathways for learners to succeed.

4.3.3 Showing more patience

"I've been using English fluently for so long that I almost forgot how hard it could be at the beginning of learning a language. It's really hard to quickly spit out a response in Chinese. So, every time I see a student struggling to respond to my questions, I give them more time and compliment on their efforts." (T1)

T1's experience shows how L3 learning can propel teachers to

transform their empathy into practice by reshaping their interactional behavior in the classroom. By adjusting her pacing and providing emotional encouragement, T1 integrates empathy into classroom interaction, creating a safer environment for learners. This illustrates how newly emerging empathy can lead to subtle but meaningful shifts in teachers' teaching practices.

4.3.4 Sharing stories of struggles with learners

"Learning something new, especially a new language can be really messy, even when you have succeeded in one language, and you have the tools and strategies you may need. That's what I told my students because that's what I experienced when learning Korean. I want to use my learning experiences when I am already a teacher of another language to inspire my students to carry on with their journey." (T5)

This extract highlights a distinctive form of empathy in which teachers transform personal learning challenges into motivational narratives for their students. She reframes difficulty as a natural and universal part of language learning rather than a sign of inadequacy. This not only normalizes students' frustrations but also reduces the psychological distance between teachers and students, which echoes the findings that NNESTs can build better rapport with students by Ma (2012). By using her own vulnerability to reassure and inspire her learners, T5 demonstrates how empathy is more than emotional understanding and becomes a pedagogical tool that supports students' persistence and confidence.

4.3.5 Learning to regulate emotions

"Feeling ashamed in the class makes me determined to make sure that no students of mine would have the same experience. I try to monitor my feelings, and make sure that I come in class with no emotional burden on me. Sometimes I journal it down, sometimes I choose to practice grounding before class. I don't want my students to be the victim of my inability to control emotions." (T4)

T4's reflection highlights an important form of empathy in which teachers translate their emotional experiences into deliberate self-

regulation practices. Her commitment reflects an awareness of the emotional influence teachers hold and the responsibility they bear in shaping a safe and supportive learning environment. In this way, T4 demonstrates that empathy is not only an internal understanding of students' feelings but also a decision to regulate one's own emotions to prevent negative consequences for learners.

In general, these extracts show a positive relationship between empathy gained from their learning experiences and their teaching practices. However, not all teachers exhibited changes after completing their L3 courses.

"I know how my students feel, but to be honest, I'm too tired to do anything. I'm under great pressure with all the teaching and paperwork, and this Japanese course..." (T6)

"Lots of work on my hand. What can I do for now? The changes may be saved for the future. I'm just trying to do what I can for now." (T3)

While many teachers enacted empathy by adjusting their teaching practices, others reported being unable to translate their insights into immediate change due to overwhelming professional demands. T6 described feeling "too tired to do anything" despite understanding her students' struggles, noting the cumulative pressure of teaching, paperwork, and her own Japanese studies. Similarly, T3 acknowledged the need for pedagogical adjustments but felt compelled to postpone them, citing the heavy workload she was managing. These accounts highlight that empathy alone might not be sufficient for practice change; it must be supported by adequate emotional, temporal, and cognitive resources. The teachers' comments reveal an important finding: Empathy can be present but not acted on when there are situational constraints, showing that heavy workloads and institutional pressures can limit even very empathetic teachers from making changes in the classroom.

The findings answer the second research question by indicating that teacher empathy contributes to EFL teaching practices by shaping how teachers provide feedback, structure learning activities, interact with students, and regulate their own emotions. Empathy encouraged teachers to prioritize learners' dignity, adjust their expectations, provide greater instructional support, and create emotionally safer learning environments. However, the findings also suggest that empathy does not always result in

changes in teaching practices, as institutional demands and workload pressures may constrain teachers' ability to enact their empathy in practice.

5. CONCLUSION

Focusing on Vietnamese NNESTs, the current study investigated the process of developing empathy through learning L3. The participants included six teachers working in four universities. The data analysis showed that all the teacher participants showed empathy towards their students with three stages: pre-existing empathy, newly emerging empathy and empathy in action. Pre-existing empathy, the empathy teachers had before starting their L3 courses, provided them with initial empathy towards their students. Newly emerging empathy is the empathy that the teachers developed during the process of L3 learning, giving them new insights into their learners' journey. Lastly, empathy in action suggests how teachers transform their inherent empathy into actionable practices. However, not all of the teachers had empathy in action due to the multiple pressures they were under.

The study highlights the importance of teacher empathy in enhancing language teaching practices in general and the role of learning a new language in fostering teacher empathy in particular. The implications of the above findings may be significant for the teaching quality and professional development of English teachers.

Firstly, the teachers' newly developed awareness about learners' cognitive and emotional challenges during their L3 learning process paved the way for the changes in their pedagogical practices such as modifying how they gave feedback to their students, exhibiting more patience, providing more scaffolding and sharing personal learning experiences with their students. This suggests that teachers should actively reflect on how their instructional choices impact their students, and they should embark on the process of learning a new foreign language to identify gaps in their teaching practice to truly understand the effects of their teaching on their students.

Secondly, teacher trainers should also encourage their trainees to gain the experience of learning a new language, and teacher training programs

should incorporate experiential learning opportunities that place teachers in the role of language learners. Such experiences can help teachers develop a deeper understanding of learners' emotional and cognitive challenges, which may not be achieved through theoretical training alone.

Moreover, the study also reveals that even when teachers had empathy towards their students, this did not immediately translate into improvements in their practices due to external constraints such as heavy workloads, paperwork, and demand from their institutions, which might mean that empathy is not only teachers' responsibility, but also institutions'. Therefore, educational institutions may need to offer more support to their teachers to enable them to implement empathetic practices.

Lastly, this research also emphasizes the value of NNESTs' multilingual experiences as a resource for their professional development. Learning the third language enabled them to develop new forms of empathy to become more aware of their students' learning experiences. This shows that NNESTs possess unique advantages over NESTs, and therefore NNESTs should be encouraged to draw on their multilingual experiences for the betterment of their teaching quality.

Considering the limited research into the development of teacher empathy in language teaching and learning, this paper provides insights for L2 teachers for the betterment of their service. This study underscores the necessity of learning a new language to gain valuable perspectives and suggests that teacher training program designers add L3 learning experience to their programs to offer professional development opportunities to their trainees. This paper might act as a suggestion for future research and development in L2 teachers' professional development, offering recommendations for enhancing L2 teaching practices.

Given the small number of participants and the specific context in which the study was conducted, the findings should be viewed as exploratory and context-specific. Rather than claiming that learning a new foreign language leads to empathy development among all NNESTs, the study provides an in-depth account of how such development was experienced by six Vietnamese EFL teachers and it is hoped to bring certain transferability to other contexts. Consequently, the findings are intended to offer insights into possible processes of empathy development rather than universally applicable conclusions.

REFERENCES

- Aldrup, Karen, Bastian Carstensen, and Uta Klusmann. 2022. Is Empathy the Key to Effective Teaching? A Systematic Review of Its Association with Teacher-Student Interactions and Student Outcomes. *Educational Psychology Review* 34:1177-1216.
- Aldrup, Karen, Bastian Carstensen, and Uta Klusmann. 2023. The role of teachers' emotion regulation in teaching effectiveness: A systematic review integrating four lines of research. *Educational Psychologist* 59(2):89-110.
- Arva, Valéria, and Péter Medgyes. 2000. Native and non-native teachers in the classroom. *System* 28(3):355-372.
- Barcelos, Ana Maria F., and Hilda Simone H. Coelho. 2016. Language learning and teaching: What's love got to do with it? In *Positive Psychology in SLA*, eds. Peter D. MacIntyre, Tammy Gregersen, and Sarah Mercer, pp.130-144. Bristol: Multilingual Matters.
- Batson, C. Daniel. 2009. These things called empathy: Eight related but distinct phenomena. In *The Social Neuroscience of Empathy*, eds. Jean Decety and William Ickes, pp.3-15. Cambridge, MA: MIT Press.
- Bolton, Kingsley. 2006. Varieties of World Englishes. In *The Handbook of World Englishes*, eds. Braj B. Kachru, Yamuna Kachru, and Cecil L. Nelson, pp.289-312. Malden, MA: Blackwell.
- Braun, Virginia, and Victoria Clarke. 2022. *Thematic Analysis: A Practical Guide*. London: Sage.
- Brown, H. Douglas, and Heekyeong Lee. 2015. *Teaching by Principles: An Interactive Approach to Language Pedagogy*. 4th edn. White Plains, NY: Pearson Education.
- Colmenero, Kebir, and David Lasagabaster. 2020. Enclosing native speakerism: Students', parents' and teachers' perceptions of language teachers. *Journal of Multilingual and Multicultural Development* 45(2):85-100.
- Cooper, Bridget. 2002. *Teachers as Moral Models: The Role of Empathy in the Relationships Between Teachers and Their Pupils*. Leeds: Leeds Metropolitan University dissertation.
- Creswell, John W., and J. David Creswell. 2018. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 5th edn. Thousand Oaks, CA: Sage.
- Curci, Antonella, Tiziana Lanciano, and Emanuela Soleti. 2014. Emotions in the classroom: The role of teachers' emotional intelligence ability in predicting students' achievement. *The American Journal of Psychology* 127(4):431-445.
- Decety, Jean, and Philip L. Jackson. 2004. The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews* 3(2):71-100.
- Derakhshan, Ali. 2021. The predictability of Turkman students' academic engagement through Persian language teachers' nonverbal immediacy and credibility. *Journal of Teaching Persian to Speakers of Other Languages* 10:3-26.
- Dewaele, Jean-Marc, and Li Wei. 2012. Multilingualism, empathy and multicompetence. *International Journal of Multilingualism* 9(4):352-366.
- Dörnyei, Zoltán. 2007. *Research Methods in Applied Linguistics*. Oxford: Oxford

- University Press.
- Ekman, Paul. 2003. *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. New York: Holt.
- Ellis, Elizabeth M. 2016. "I may be a native speaker but I'm not monolingual": Reimagining all teachers' linguistic identities in TESOL. *TESOL Quarterly* 50(3):597-630.
- Ellis, Rod. 2021. Explicit and implicit oral corrective feedback. In *The Cambridge Handbook of Corrective Feedback in Second Language Learning and Teaching*, eds. Hossein Nassaji and Eva Kartchava, pp.341-364. Cambridge: Cambridge University Press.
- Franklin, Melissa M. 2013. *Teacher Impact on the Academic Achievement of Students of Poverty*. California: TUI University dissertation.
- Gkonou, Christina, and Sarah Mercer. 2017. *Understanding Emotional and Social Intelligence among English Language Teachers*. British Council ELT Research Papers.
- Goroshit, Marina, and Meriav Hen. 2016. Teachers' empathy: Can it be predicted by self-efficacy? *Teachers and Teaching: Theory and Practice* 22(7):805-818.
- Guiora, Alexander Z., Robert C. L. Brannon, and Cecelia Y. Dull. 1972. Empathy and second language learning. *Language Learning* 22(1):111-130.
- Hashim, Nik Mohd Hazrul Nik, Syed Shah Alam, and Norazlina Mohd Yusoff. 2014. Relationship between teacher's personality, monitoring, learning environment, and students' EFL performance. *GEMA Online Journal of Language Studies* 14(1):101-116.
- Hieu, Hua Hong, and Le Thanh Thao. 2024. Integrating empathy in EFL education: A study of Vietnamese tertiary teachers' adaptive responses. *Journal of Language Teaching and Research* 15(4):1381-1389.
- Hodges, Sara D., and Michael W. Myers. 2007. Empathy. In *Encyclopedia of Social Psychology*, Vol. 1, eds. Roy F. Baumeister and Kathleen D. Vohs, pp.296-298. Thousand Oaks, CA: Sage.
- Howe, David. 2013. *Empathy: What It Is and Why It Matters*. Basingstoke: Palgrave Macmillan.
- Johnson, R. Burke, and Larry B. Christensen. 2016. *Educational Research: Quantitative, Qualitative, and Mixed Approaches*. 6th edn. Thousand Oaks, CA: Sage.
- Klassen, Robert, Tracy L. Durksen, Lisa Kim, Fiona Patterson, Emma Rowett, Jane Warwick, Paul Warwick, and Mary Wolpert. 2017. Developing a proof-of-concept selection test for entry into primary teacher education programs. *International Journal of Assessment Tools in Education* 4(2):96-114.
- Latchaw, Jody. 2016. *Online Postsecondary Student Perception of Instructor Emotional Intelligence and Student Performance: A Quantitative Correlational Study*. Arizona: Northcentral University dissertation.
- Le, Thanh Thao, Trut Thuy Pham, and Thi Thuy Tran. 2024. Empathy in practice: Exploring EFL teachers' strategies within the cultural landscape of Vietnam's Mekong Delta higher education institutions. *European Journal of English Language Studies* 4(2):67-79.
- Linville, Heather A. 2015. ESOL teachers as advocates: An important role? *TESOL*

- Journal* 7(1):98-131.
- Lipovsky, Caroline, and Ahmar Mahboob. 2010. Appraisal of native and non-native English-speaking teachers. In *The NNEST Lens: Non-Native English Speakers in TESOL*, ed. Ahmar Mahboob, pp.154-179. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Ma, La Ping Florence. 2012. Advantages and disadvantages of native- and nonnative-English-speaking teachers: Student perceptions in Hong Kong. *TESOL Quarterly* 46(2):280-305.
- Maxwell, Bruce. 2017. Pursuing the aim of compassionate empathy in higher education. In *The Pedagogy of Compassion at the Heart of Higher Education*, ed. Paul Gibbs, pp.33-48. Cham: Springer.
- McNeill, Arthur. 2005. Non-native speaker teachers and awareness of lexical difficulty in pedagogical texts. In *Non-Native Language Teachers: Perceptions, Challenges, and Contributions to the Profession*, ed. Enric Llurda, pp.107-128. New York: Springer.
- Medgyes, Péter. 1994. *The Non-Native Teacher*. London: Macmillan.
- Mercer, Sarah. 2016. Seeing the world through your students' eyes: Empathy in language learning and teaching. In *Positive Psychology in SLA*, eds. Peter D. MacIntyre, Tammy Gregersen, and Sarah Mercer, pp.91-111. Bristol: Multilingual Matters.
- Meyers, Sal, Katherine Rowell, Mary Wells, and Brian C. Smith. 2019. Teacher empathy: A model of empathy for teaching for student success. *College Teaching* 67(3):160-168.
- Moodie, Ian. 2016. The anti-apprenticeship of observation: How negative prior language learning experience influences English language teachers' beliefs and practices. *System* 60:29-41.
- Mullock, Barbara. 2010. Does a good language teacher have to be a native speaker? In *The NNEST Lens: Non-Native English Speakers in TESOL*, ed. Ahmar Mahboob, pp.87-113. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Nemtchinova, Ekaterina. 2010. The "who's worth more?" question revisited: MA TESOL practicum host teachers' perceptions of NES and NNEST teacher trainees. In *The NNEST Lens: Non-Native English Speakers in TESOL*, ed. Ahmar Mahboob, pp.129-153. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Ocampo Jaramillo, Marcela Angelita, Ana María Solano, and Joseph Fritzner. 2025. Empathy in language teaching in higher education. *Ciencia Latina Revista Científica Multidisciplinar* 9(4):1801-1815.
- Oliver, Rhonda, and Rebecca Adams. 2021. Oral corrective feedback. In *The Cambridge Handbook of Corrective Feedback in Second Language Learning and Teaching*, eds. Hossein Nassaji and Eva Kartchava, pp.187-206. Cambridge: Cambridge University Press.
- Oxford, Rebecca L., and Rashi Jain. 2010. Students' evolving perspectives on World Englishes, non-native English speakers, and non-native English-speaking teachers based on a graduate course. In *The NNEST Lens: Non-Native English Speakers in TESOL*, ed. Ahmar Mahboob, pp.239-262. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Ozanska-Ponikwia, and Jean-Marc Dewaele. 2012. Personality and L2 use: Open-minded

- and self-confident individuals are more likely to be active and proficient L2 users. *EUROSLA Yearbook* 12:112-134.
- Parchomiuk, Monika. 2019. Teacher empathy and attitudes towards individuals with disabilities. *International Journal of Disability, Development and Education* 66(1):56–69.
- Pidbutska, Nina, Yuliia Demidova, and Anastasiia Knysh. 2021. Gender aspects of empathy in online learning of adolescents. *Journal of Education, Culture and Society* 12(1):314–321.
- Roorda, Debora L., Suzanne Jak, Margaretha Zee, Frans J. Oort, and Helma M. Y. Koomen. 2017. Affective teacher-student relationships and students' engagement and achievement: A meta-analytic update and test of the mediating role of engagement. *School Psychology Review* 46(3):239-261.
- Roothoof, Hanne, and Ruth Breeze. 2016. A comparison of EFL teachers' and students' attitudes to oral corrective feedback. *Language Awareness* 25(4):318-335.
- Rota, Giuseppina, and Susanne Maria Reiterer. 2009. Cognitive aspects of pronunciation talent: How empathy, mental flexibility, working memory and intelligence interact with phonetic talent. In *Language Talent and Brain Activity*, eds. Grzegorz Dogil and Susanne Maria Reiterer, pp.67-96. Berlin: Mouton de Gruyter.
- Sadoughi, Majid, and S. Yahya Hejazi. 2021. Teacher support and academic engagement among EFL learners: The role of positive academic emotions. *Studies in Educational Evaluation* 70:101060.
- Salem, Mansoureh, and Omid Tabatabaei. 2015. The relationship between emotional intelligence and emotional empathy of EFL teachers regarding gender. *International Journal of Educational Investigations* 2(8):146-154.
- Segal, Elizabeth A. 2011. Social empathy: A model built on empathy, contextual understanding, and social responsibility that promotes social justice. *Journal of Social Service Research* 37(3):266-277.
- Sharifian, Farzad. 2009. *English as an International Language: Perspectives and Pedagogical Issues*. Bristol: Multilingual Matters.
- Smith, Adam. 2006. Cognitive empathy and emotional empathy in human behavior and evolution. *The Psychological Record* 56(1):3-21.
- Tettegah, Sharon, and Carolyn J. Anderson. 2007. Pre-service teachers' empathy and cognitions: Statistical analysis of text data by graphical models. *Contemporary Educational Psychology* 32(1):48-82.
- Teo, Timothy, Saeed Khazaie, and Ali Derakhshan. 2022. Exploring teacher immediacy-(non)dependency in the tutored augmented reality game-assisted flipped classrooms of English for medical purposes comprehension among Asian students. *Computers & Education* 179:104406.
- Walkinshaw, Ian, and Duongthi Hoang Oanh. 2014. Native and non-native English language teachers: Student perceptions in Vietnam and Japan. *SAGE Open* 4(2).
- Willis, Rohan K., Sharon Harvey, and Irmengard Wohlfart. 2025. Enhanced empathy for English language learners: How teachers' past language learning informs their teacher knowledge and practices. *TESOL Quarterly* 59(4):2061–2092.

- Wink, Mackenzie N., Maria D. LaRusso, and Rhiannon L. Smith. 2021. Teacher empathy and students with problem behaviors: Examining teachers' perceptions, responses, relationships, and burnout. *Psychology in the Schools* 58(8):1575-1596.
- Zhang, Zhichao. 2022. Toward the role of teacher empathy in students' engagement in English language classes. *Frontiers in Psychology* 13:880935.
- Zhu, Juncheng, Xin Qiang Wang, Xiaoxin He, Yuan-Yu Hu, Fuhong Li, Ming-Fan Liu, and Baojuan Ye. 2019. Affective and cognitive empathy in pre-teachers with strong or weak professional identity: An ERP study. *Frontiers in Human Neuroscience* 13:175.

Linh Thuy Nguyen
International Education Center
Hanoi University
Dai Mo Ward, Hanoi, Viet Nam
linhthuy.work210@gmail.com

透過學習新外語發展非母語英語教師之同理心

阮翠玲
河內大學

教師同理心有助於營造正向學習環境及提升教學品質，但英語語言教學領域對教師同理心及其發展仍缺乏探討。本研究採用質性研究設計，探討沉浸式學習新外語如何影響語言教師的同理理解與教學實踐。研究對象為越南四所大學六位非母語英語教師，透過焦點團體訪談、學習日誌、半結構式訪談及主題分析法進行分析。結果顯示，學習新外語有助於重新喚起並發展教師對學生的同理心，促使其調整教學實踐；持續學習外語亦能培養非母語英語教師的動態同理心，提升教學成效。

關鍵詞： 教師同理心、社會情緒能力、情緒智力、非母語英語教師

LECTURERS' BELIEFS ABOUT NATIONAL LANGUAGE POLICIES IN ENGLISH MEDIUM INSTRUCTION PROGRAMS IN VIETNAM

Tran Tan Phat^{1,2}

¹*Ho Chi Minh University of Banking*

²*Ho Chi Minh City Open University*

ABSTRACT

The research investigates lecturers' beliefs about national language policies in English medium instruction (EMI) programs in Vietnam, with a focus on how these beliefs shape classroom practices. Grounded in a qualitative research design, the research was conducted at Ho Chi Minh University of Banking and involved ten EMI lecturers across disciplines. The data were collected through questionnaires, classroom observations, and interviews to capture both stated beliefs and actual teaching practices. The findings reveal that lecturers hold positive attitudes toward EMI, perceiving it as a pathway to enhanced access to international knowledge and improved student employability. However, these beliefs are mediated by contextual factors, particularly students' English proficiency, institutional support, and disciplinary demands. Lecturers reported moderate but persistent challenges in implementing EMI policies, including increased workload, difficulties balancing content and language instruction, and the absence of clear policy guidelines. While national policies tend to promote English-only instruction, lecturers frequently adopt flexible language practices, including the strategic use of Vietnamese, to support student comprehension. This reflects a shift toward a more functional and context-sensitive understanding of language use in EMI settings.

Keywords: English medium instruction (EMI), Ho Chi Minh University of Banking, Language policies, Language proficiency, Lecturers' beliefs

1. INTRODUCTION

The internationalisation of higher education has profoundly reshaped language policies across the globe, with English as a Medium of Instruction (EMI) emerging as a dominant force in non-native English-speaking countries (Dearden 2015; Macaro and Searle 2025). Vietnam represents a compelling context within this global phenomenon, where EMI has been implemented since the 1990s as part of a broader neoliberal agenda to modernise the economy, enhance graduate employability, and integrate into the global knowledge economy (Vu and Burns 2014; Nguyen et al. 2017). The Vietnamese government's National Foreign Languages 2020 Project, later extended to 2025, established ambitious goals for educational reform, positioning EMI as a strategic instrument to achieve both educational and non-educational objectives in the global age (Nguyen 2022; Tran et al. 2026).

National language policies in Vietnam have encouraged – though not mandated – EMI implementation in selected higher education institutions meeting specific government requirements (Nguyen 2018; Pham and Doan 2020). These policies are underpinned by ideological assumptions that frame EMI as a gateway to linguistic, cultural, economic, and social capital (Bourdieu 1991; Nguyen et al. 2024), with the potential to advance national socio-economic development and individual career prospects (Tran et al. 2022). However, a critical examination of policy documents reveals notable absences: explicit discussions regarding implementation challenges, quality assurance mechanisms, and the practical realities of EMI delivery remain largely unaddressed at the policy level (Malmström and Zhou 2025; Nguyen 2022; Pham and Doan 2020).

Within this policy landscape, lecturers occupy a pivotal yet often underexamined position. They function as “primary policymakers” at the grassroots level, exercising agency to interpret, adapt, and implement top-down language policies according to their classroom realities (Nguyen and Nguyen 2020). The beliefs that lecturers hold about national language

policies are therefore consequential, shaping how policies are translated into pedagogical practice and ultimately determining the success or failure of EMI initiatives (Borg 2006; Nguyen and Nguyen 2020). Recent scholarship has highlighted that lecturers' beliefs encompass complex dimensions: their perceptions of EMI benefits and challenges, their views on language use and translanguaging practices, and their responses to the ideological tensions embedded within English-only policy mandates (Rose et al. 2020; Sahan and Şahan 2022; Pham et al. 2023).

Research conducted between 2020 and 2025 reveals that Vietnamese lecturers generally maintain positive attitudes towards EMI, recognising its potential to enhance students' access to international knowledge and improve employment outcomes (Nguyen and Nguyen 2020; Pham et al. 2023). Yet this positivity coexists with significant challenges. Lecturers report substantial workload increases, concerns about students' insufficient English proficiency, and the pedagogical complexities of delivering content through a second language (Tran et al. 2022). Moreover, stark misalignments have been documented between the "English-only" ideologies explicit in national and institutional policies and the actual language practices occurring in classrooms, where lecturers and students strategically employ Vietnamese as a pedagogical tool.

A growing body of literature positions lecturers not as passive policy recipients but as active "change agents" who exercise professional agency to reconcile policy expectations with student needs (Nguyen and Bui 2023; Tran and Nguyen 2018). This reconceptualization challenges traditional top-down models of language policy implementation and calls for greater attention to the beliefs, practices, and agency of those charged with enacting policy at the micro-level. As Sahan and Şahan (2022) argue, effective EMI implementation requires consultation with key stakeholders and careful consideration of local contexts, rather than the uncritical adoption of English-only approaches that reflect native-speaker ideologies.

2. LITERATURE REVIEW

Research on lecturers' beliefs in EMI contexts draws on several theoretical traditions that illuminate the complex relationships between

policy, ideology, and practice. Spolsky's (1998) tripartite framework of language policy – comprising language management, language beliefs, and language practices – has proven particularly influential in Vietnamese EMI research (Nguyen and Nguyen 2020; Pham et al. 2023). This framework enables researchers to examine how national policies (management) interact with lecturers' ideological orientations (beliefs) and actual classroom behaviours (practices). Ricento and Hornberger's (1996) "onion" model of language policy layers has similarly informed studies that position lecturers as pivotal actors mediating between macro-level policy and micro-level classroom realities (Rose et al. 2020).

The concept of language ideology has emerged as a critical lens for understanding how EMI policies are both constructed and contested (Nguyen et al. 2017). Drawing on this theoretical orientation, Tri and Moskovsky (2023) conducted a multi-layered examination of language ideologies in Vietnamese EMI education, analysing policy documents and stakeholder discourse across national and institutional levels. His findings revealed that EMI is ideologically constructed as producing multiple forms of capital – linguistic, cultural, economic, and social – with enormous potential for national socio-economic development and individual employability (Tran et al. 2022). This ideological framing, however, carries unintended consequences: the celebration of EMI as access to "supreme quality education" grants it symbolic power that simultaneously downgrades the status of Vietnamese as a medium of instruction (Nguyen 2022; Pham et al. 2023).

More recently, ecological approaches to agency have informed understandings of how lecturers navigate these ideological tensions. Pham et al. (2023) employed an ecological model to investigate tertiary academic agency in Vietnamese EMI programs, demonstrating that agency emerges not as an inherent capacity but through dynamic interactions between individual capabilities and environmental factors (Biesta and Tedder 2006; Nguyen and Nguyen 2020). Her research identified three interrelated types of agency enacted by lecturers: pedagogical agency (capacity to make informed teaching decisions), relational agency (ability to collaborate and build professional relationships), and reflexive agency (capacity for critical self-reflection and adaptation). This ecological perspective shifts analytical focus from

viewing lecturers as passive policy recipients to recognising them as active agents whose beliefs shape implementation outcomes (Priestley et al. 2015; Tran and Nguyen 2018).

2.1 The Policy Context: Ideological Foundations and Gaps

National language policies in Vietnam have established the framework within which EMI programs operate, yet significant gaps exist between policy rhetoric and implementation realities. Tran et al.'s (2022) research analysed national policy documents, revealing that while EMI is positioned as a strategic instrument for achieving educational and economic goals, explicit discussions of implementation challenges and quality assurance remain limited (Nguyen 2022). This aligns with broader regional findings that EMI policies often prioritise internationalisation goals over pedagogical feasibility (Macaro and Searle 2025).

The National Foreign Languages 2020 Project represents the government's flagship initiative for language reform. Tri and Moskovsky (2023) highlighted that despite increased institutional autonomy, translating policy into practice remains complex, often resulting in an "implementation gap" (Nguyen and Nguyen 2020; Pham et al. 2023; Tollefson and Tsui 2014). Similar tensions have been observed across Asian EMI contexts, where policy aspirations frequently exceed institutional capacity (Rose et al. 2020). According to Rose et al.'s perspectives, they provided a critical review of EMI policies, identifying misalignments between policy and practice and questioning the monolingual ideology underpinning EMI. Rose's argument resonates with global critiques of English-only approaches, which have been challenged for marginalising local languages and epistemologies (Pennycook 2017).

2.2 Lecturers' Beliefs: Perceptions of EMI Benefits

Research consistently demonstrates that Vietnamese lecturers hold predominantly positive attitudes towards EMI (Nguyen et al. 2024; Pham et al. 2023). Tran et al.'s (2022) study found that lecturers perceive EMI as enhancing access to international knowledge and improving employability outcomes. These findings align with global EMI research,

which highlights perceived benefits in academic mobility and knowledge access (Dearden 2015; Macaro and Aizawa 2022). Tri and Moskovsky (2023) found a shared “English-only” ideology across stakeholders, reinforcing EMI’s perceived value in internationalisation.

2.3 Perceived Challenges and Barriers

Despite positive perceptions, lecturers report significant challenges. Tran et al. (2022) identified students’ low English proficiency and increased workload as major barriers. These findings are widely echoed in EMI literature across Asia (Kirkpatrick 2017; Macaro 2018; Rose et al. 2020). They identified systemic issues, including limited teacher preparation, inequitable access, and reliance on Western-centric materials. Such challenges raise concerns about educational quality and equity. Pham et al. (2023) demonstrated that institutional constraints – such as performativity pressures and rigid communication structures – limit lecturer agency, while professional networks and collaboration enhance it. These findings align with ecological models emphasising context-dependent agency (Nguyen et al. 2024; Priestley et al. 2015).

2.4 Language Practices: Between Policy and Reality

A key finding across studies is the misalignment between English-only policies and classroom practices. Tri and Moskovsky (2023) found that lecturers frequently employ Vietnamese strategically, despite policy expectations. This reflects broader evidence that translanguaging is a pragmatic and effective pedagogical strategy in EMI contexts (Sahan and Şahan 2022; Zheng and Choi 2024). They confirmed that while stakeholders express preferences for English-only instruction, the first language remains an essential classroom resource. Similarly, Pham et al. (2023) highlighted the co-construction of EMI practices by teachers and students through adaptive strategies. The increasing recognition of translanguaging reflects a paradigm shift in EMI research, challenging monolingual ideologies and advocating for multilingual pedagogies. This perspective underscores the need for policy frameworks that align more closely with classroom realities and support flexible language use.

3. METHODOLOGY

Building on the theoretical and contextual foundations outlined above, the research seeks to explore the complex interplay between language policy, lecturer cognition, and classroom practice within EMI programmes in Vietnam. In particular, it addresses four interrelated research questions that guide the investigation:

1. What are lecturers' beliefs about national language policies in EMI programs?
2. How do contextual factors such as student proficiency and institutional support influence lecturers' policy beliefs?
3. What challenges do lecturers identify as most significant in implementing national EMI policies?
4. In what ways do lecturers' beliefs about language use differ from the expectations articulated in policy documents?

To address these research questions, the study adopts a qualitative research design, which is particularly well suited to examining complex, context-dependent phenomena such as language beliefs and pedagogical practices. As noted by Creswell and Creswell (2018), qualitative inquiry enables researchers to explore how individuals interpret their experiences within specific social and institutional contexts. In EMI settings, where language use is deeply intertwined with cultural, ideological, and pedagogical considerations, such an approach is essential. Rather than reducing lecturers' perceptions to quantifiable variables, qualitative methods prioritise depth, nuance, and meaning, thereby capturing the multifaceted realities of teaching through English. This perspective is further supported by Dörnyei (2007), who emphasises that qualitative research facilitates a richer understanding of participants' motivations, beliefs, and decision-making processes.

The research was conducted at Ho Chi Minh University of Banking (HUB), a prominent public institution in Vietnam specialising in economics, finance, business administration, etc. Established on 12/16/1976 and operating under the State Bank of Vietnam, HUB has developed into a key centre for higher education, serving a large

population of undergraduate and postgraduate students across multiple faculties. The university offers a diverse range of programmes, including undergraduate, master's, and doctoral degrees in fields such as finance, banking, business administration, and education. In recent years, HUB has increasingly prioritised internationalisation as a strategic objective, integrating EMI programmes into its curriculum to prepare students for participation in global academic and professional environments.

A central component of this internationalisation effort is the SaigonISB International Training Institute (SaigonISB), an affiliated unit responsible for delivering internationally oriented programmes. Through partnerships with reputable universities from countries with well-established education systems, SaigonISB provides students with access to globally recognised curricula while remaining within the Vietnamese educational context. EMI programmes at SaigonISB aim not only to enhance students' English proficiency but also to develop their academic literacy and intercultural competence. In doing so, the institute plays a pivotal role in aligning HUB's educational offerings with international standards and labour market demands.

The participants in the research consisted of ten lecturers teaching EMI courses ranging from A1 to B1 levels (see Table 1). These lecturers were selected to provide insight into a range of teaching experiences and perspectives within EMI classrooms.

Data collection was conducted in multiple stages to ensure a comprehensive and triangulated understanding of lecturers' beliefs and practices. Initially, all participants completed an open-ended questionnaire designed to elicit their views on English language use, EMI policy, and pedagogical approaches. This phase established a foundational understanding of lecturers' beliefs prior to direct classroom observation.

Subsequently, classroom observations were conducted over two to three sessions per a lecturer, with each session lasting between two and three hours. This extended observation period allowed the researchers to examine instructional strategies, interaction patterns, and language use in authentic teaching contexts. Observations were particularly valuable in identifying potential discrepancies between lecturers' stated beliefs and their actual classroom practices. Finally, five lecturers agreed to participate in follow-up interviews, which provided deeper insight into

their experiences, challenges, and reflections on EMI implementation. These interviews enabled the researchers to explore not only what lecturers do, but also why they make particular pedagogical choices.

Table 1. List of Participants: EMI Lecturer

No	Teaching experience		EMI major	Interview
	EMI years	Total years		
L1	3	8	Business Administration	-
L2	7	12	Business Administration	x
L3	1	3	Supply Chain Management	x
L4	0.6	2	Supply Chain Management	-
L5	12	19	Marketing	x
L6	3	4	Marketing	-
L7	1	3	Finance	-
L8	13	15	Finance	x
L9	8	11	Finance – Banking – Insurance	x
L10	5	7	Finance – Banking – Insurance	-

4. RESULTS AND DISCUSSION

4.1 Lecturers' Beliefs about National Language Policies and Their Implementation in EMI Programs

Lecturers' beliefs about national language policies in English Medium Instruction (EMI) programs reveal a complex negotiation between policy expectations, pedagogical realities, and the evolving global status of English. While national policies often promote standardized forms of English aligned with native-speaker norms, many lecturers appear to adopt a more pragmatic and flexible stance in their actual teaching practices. The findings reflected in Tables 2 and 3 suggest that lecturers prioritize communicative effectiveness – evidenced by the unanimous emphasis on confidence (10/10) and high attention to pronunciation (7/10) and fluency (6/10) – over strict adherence to linguistic accuracy (5/10). This pattern indicates a belief that successful communication in EMI contexts is not solely dependent on grammatical precision, but rather on the ability to

convey meaning clearly and confidently in diverse academic settings. Similarly, in writing assessment, lecturers place considerable importance on organization (9/10) and content (7/10), which further underscores their orientation toward meaning-making and discourse competence rather than purely formal correctness. These tendencies suggest that lecturers may perceive national language policies as somewhat idealized or disconnected from the multilingual realities of EMI classrooms, where students often possess varying levels of English proficiency and rely on English as a lingua franca rather than a native language.

Table 2. Summary of Lecturers' Criteria to Assess Speaking Ability

Category	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	Total (people)
Pronunciation	x	x	x	-	x	x	-	-	x	x	7/10
Accuracy	x	x	x	x	-	x	-	-	-	-	5/10
Confident	x	x	x	x	x	x	x	x	x	x	10/10
Fluency	x	-	x	x	x	x	-	-	x	-	6/10
Contents	x	x	x	x	-	-	-	x	x	x	7/10

Table 3. Summary of Lecturers' Criteria to Assess Writing Ability

Category	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	Total (people)
Spelling	x	x	x	x	x	x	-	-	x	x	8/10
Grammar	x	x	x	x	-	-	-	x	x	x	7/10
Creation	x	x	x	x	-	x	x	x	-	-	7/10
Organization	x	-	x	x	x	x	x	x	x	x	9/10

Consequently, lecturers seem to reinterpret policy guidelines through a contextual lens, adapting them to better support students' academic success and participation. This aligns with the argument that English today functions as a global communicative resource rather than the exclusive property of native speakers, thereby necessitating a shift from prescriptive norms to more inclusive and functional approaches to language use. In this regard, lecturers' beliefs appear to resonate with contemporary

perspectives that advocate for the integration of diverse English varieties and the recognition of linguistic plurality in teaching practices. By valuing confidence, intelligibility, and organizational clarity, lecturers implicitly challenge rigid policy frameworks that prioritize native-like accuracy, instead endorsing a more dynamic and context-sensitive understanding of language competence. However, this divergence between policy and practice may also reflect underlying tensions, as lecturers must navigate institutional expectations while addressing students' immediate learning needs. Factors such as limited institutional support, heterogeneous student proficiency, and the pressure to meet international academic standards further complicate this alignment. As a result, lecturers' beliefs can be seen as both adaptive and critical, simultaneously accommodating policy objectives and reshaping them to fit local contexts. Ultimately, these findings suggest that effective EMI implementation requires a more nuanced and flexible policy approach – one that acknowledges the realities of global English use and empowers lecturers to exercise professional judgment. By bridging between policy ideals and classroom practices, lecturers play a pivotal role in transforming EMI into a more inclusive, realistic, and pedagogically sound educational model that prepares students for participation in global academic and professional communities.

Lecturers in this research demonstrated similar attitudes toward students' pronunciation in both their teaching practices and English proficiency test requirements. Consistent with the findings of Dang and Cross (2023), most lecturers acknowledge that Vietnamese students tend to use English primarily for academic purposes. As a result, they exhibit a certain level of acceptance toward this linguistic behavior. However, despite this awareness, there remains a strong preference for students to adhere to native-speaker norms when preparing for standardized English proficiency tests. This is largely due to the fact that major English tests, such as TOEFL and IELTS, continue to be guided by first-language English norms (Nguyen and Nguyen 2020). Consequently, lecturers in this study opposed the use of non-native English varieties in formal assessments. Their reasoning aligns with the perception that English proficiency tests are inherently “formal” contexts, where Vietnamese English varieties are deemed inappropriate, a perspective explicitly

mentioned by L2, L3, L6, and L7. Furthermore, the lecturers' assessment criteria predominantly focused on enhancing students' linguistic competence, particularly in terms of accuracy and pronunciation. They also placed considerable emphasis on attitudinal factors, such as confidence, when evaluating speaking performance. However, an important gap in their assessment approach was observed, as none of the lecturers took cultural aspects into account when assessing students' English communication skills. This suggests that while linguistic accuracy and fluency were prioritized, the role of English as a global language, which encompasses diverse cultural expressions and pragmatic uses, was largely overlooked. The findings highlight a discrepancy between the lecturers' recognition of students' academic English use and their insistence on first-language norms for test performance, underscoring the ongoing influence of native-speaker ideologies in English language teaching and assessment.

4.2 Contextual Factors Influence Lecturers' Policy Beliefs

The data in Figure 1 indicate that lecturers' beliefs about national language policies in EMI programs are significantly shaped by a range of contextual factors, which interact in nuanced and discipline-specific ways. Among these, individual and social-cultural factors consistently receive the highest mean scores across most fields, suggesting that lecturers perceive personal teaching experience, pedagogical beliefs, and the broader sociocultural environment as particularly influential in shaping their interpretations of EMI policies. For instance, in Marketing, individual factors reach a peak mean of 3.22, while social and cultural factors closely follow at 3.21, highlighting the extent to which lecturers' professional identities and students' cultural backgrounds inform their policy-related decisions. This pattern implies that EMI implementation is not merely a technical or linguistic exercise but a deeply situated practice embedded in specific educational contexts. Similarly, institutional support factors, with means ranging from 2.87 to 3.00, play a crucial role, reflecting the importance of administrative guidance, training opportunities, and resource availability in enabling lecturers to align their practices with policy expectations. In contrast, student proficiency factors,

although still relevant, tend to receive slightly lower scores, particularly in Marketing (2.78), which may indicate that lecturers view linguistic competence as only one component within a broader constellation of influences. Mental factors, including stress and cognitive load, also demonstrate a moderate impact across all disciplines, underscoring the psychological demands associated with teaching complex content in a second language. Notably, variations across disciplines – such as the relatively higher emphasis on social-cultural factors in Business Administration (3.12) and the stronger influence of mental factors in Supply Chain Management (3.05) – suggest that lecturers’ beliefs are further mediated by the epistemological characteristics and communicative demands of different academic fields.

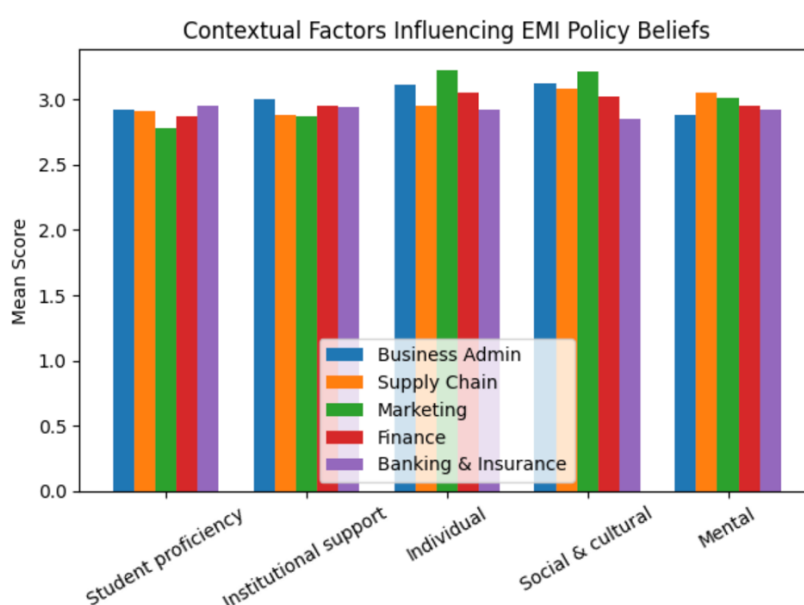


Figure 1. Compare mean scores across contextual factors

Overall, these findings illustrate that lecturers do not interpret national EMI policies in isolation; rather, their beliefs are dynamically constructed through the interplay of personal, institutional, and sociocultural contexts.

This underscores the need for flexible and context-sensitive policy frameworks that acknowledge disciplinary diversity and support lecturers in navigating the complex realities of EMI implementation.

4.3 Lecturers Identify Challenges in Implementing National EMI Policies

The data presented in Figure 2 highlight that lecturers perceive the implementation of national language policies in EMI programs as a moderately challenging endeavor, shaped by a combination of linguistic, pedagogical, and institutional constraints. A clear pattern emerges in which the majority of responses cluster around the “intermediate” level, with percentages ranging from 40% to 60% across all lecturers. This suggests that, while EMI policies are not viewed as overwhelmingly difficult, they nonetheless pose persistent and multifaceted challenges in day-to-day teaching practice. One key issue underlying this perception is the disparity in students’ English proficiency, which compels lecturers to constantly adjust their instructional strategies. For instance, when students struggle to comprehend academic content delivered in English, lecturers may need to simplify language, repeat explanations, or even resort to code-switching, thereby deviating from policy expectations that often emphasize exclusive English use. Furthermore, the relatively lower but still notable percentages in the “difficult” and “extremely difficult” categories indicate that certain aspects of EMI implementation – such as assessing students fairly or maintaining academic rigor – can become particularly demanding under constrained conditions. Lecturers who reported higher levels of difficulty may be those working in contexts with limited institutional support, including insufficient training, lack of teaching resources, or unclear policy guidelines. In contrast, the smaller proportion of “easy” responses, generally ranging between 10% and 20%, implies that only a minority of lecturers feel fully confident in navigating EMI requirements without significant obstacles. This imbalance underscores the gap between policy ideals and classroom realities, where lecturers must reconcile top-down mandates with the practical needs of diverse learner populations.

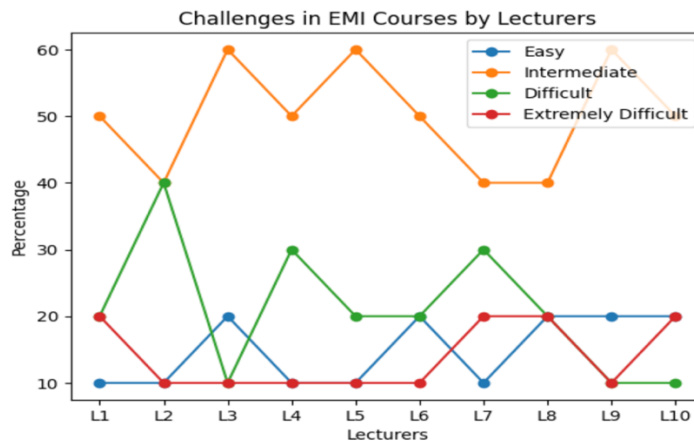


Figure 2. The challenges of the EMI courses

Additionally, the variation across individual lecturers suggests that perceptions of difficulty are not uniform but are instead influenced by contextual factors such as teaching experience, disciplinary background, and access to professional development. Importantly, the predominance of “intermediate” responses can also be interpreted as evidence of lecturers’ adaptive capacity: rather than rejecting EMI policies outright, they engage with them critically and pragmatically, finding workable solutions despite ongoing challenges. This adaptive stance reflects a broader understanding of EMI as a dynamic and evolving practice rather than a fixed instructional model. Ultimately, the findings indicate that practical implementation of national EMI policies depends not only on the clarity and coherence of the policies themselves but also on the extent to which lecturers are supported in addressing the inherent complexities of teaching content through a second language.

4.4 Lecturers’ Beliefs of Language Use

The data in Figure 3 provide valuable insights into lecturers’ beliefs about language use in EMI programs, revealing a clear prioritization of certain pedagogical functions over others. Most notably, language style

emerges as the dominant concern, accounting for approximately half of the total responses (50.5%), which suggests that lecturers place considerable emphasis on how language is delivered in the classroom. This focus reflects a belief that clarity, appropriateness of register, and adaptability of linguistic choices are essential for facilitating students' comprehension of disciplinary content in English. In contrast, communication strategies represent a smaller yet still significant proportion (19.2%), indicating that lecturers recognize the importance of techniques such as paraphrasing, repetition, and code-switching to support meaning-making, particularly in linguistically diverse classrooms. Feedback (18.2%) is also moderately emphasized, suggesting that lecturers value the role of corrective and formative responses in guiding students' language development, although it may not be as central as real-time instructional delivery. Meanwhile, explanation receives the least attention (12.1%), which may imply that lecturers assume explanation is inherently embedded within other practices, such as language style and communication strategies, rather than functioning as a distinct category.

Overall, these patterns suggest that lecturers' beliefs about language use are largely oriented toward ensuring intelligibility and accessibility rather than strictly adhering to prescriptive norms of accuracy. This aligns with a more functional view of English in EMI contexts, where the primary goal is effective knowledge transmission rather than linguistic perfection. Furthermore, the relatively balanced distribution across categories – aside from the prominence of language style – indicates that lecturers adopt a multifaceted approach to language use, integrating various strategies to address the complex demands of teaching content through a second language. Such findings underscore the need for EMI policies to acknowledge the practical realities of classroom interaction and to support lecturers in developing flexible, context-sensitive language practices.

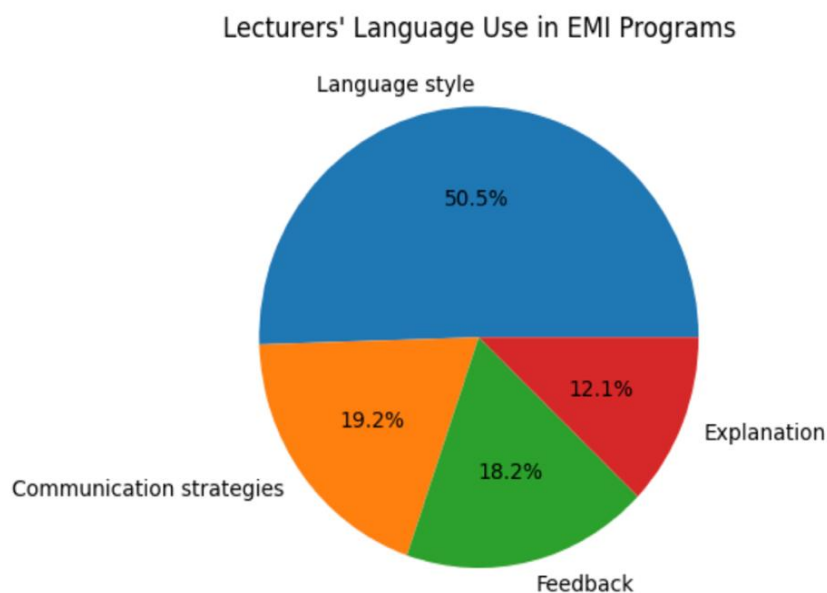


Figure 3. Summary of Lecturers' Language Use

4.5 Discussion

According to the research, lecturers had no influence over the choice of instructional materials because these choices were determined at the institutional level. This is consistent with the findings of Walsh (2010), who similarly discovered that pupils were mostly exposed to native English types through pre-recorded tapes and that teachers lacked control in selecting instructional content. In a similar vein, interviewees in this study said that they had little control over the choice of textbooks used in their lectures when they were employed by educational institutions, universities, or English centers. Teachers were forced to adhere to the specified curriculum since the board of management made these decisions instead. The dominance of native-speaker norms in EMI classrooms is reinforced by this lack of control over instructional materials, which also

limits instructors' ability to adapt lessons to students' requirements. The research's inability to investigate the board of management's viewpoints and awareness of English teaching procedures is a major disadvantage due to this institutional constraint. Without insights from decision-makers at the management level, it remains unclear why specific materials are chosen and whether considerations beyond linguistic accuracy and fluency are taken into account. Understanding the rationale behind textbook selection would provide a more comprehensive view of how English is taught in EMI settings and whether there is room for greater inclusivity of diverse English varieties.

Future research should investigate the attitudes of educational policymakers and administrators toward English language teaching materials to determine whether their decisions align with contemporary perspectives on English as a global language. This would help bridge the gap between institutional policies and lecturers' pedagogical autonomy, ensuring a more effective and contextually appropriate approach to EMI instruction.

5. CONCLUSION

5.1 Concluding Sentence

Although lecturers in this research demonstrated a general awareness of English teaching practices in EMI settings, they lacked a clear conceptual understanding of the term. As a result, they struggled to effectively incorporate EMI principles into their teaching. This aligns with previous research, which suggests that while lecturers acknowledge the role of English as a global language, their pedagogical approaches remain influenced by traditional teaching models (Nguyen and Nguyen 2020; Walsh 2010). Additionally, lecturers in this research expressed strong expectations for students to use native English varieties in high-stakes English proficiency tests, a reflection of the ongoing dominance of native-speaker norms in language assessment (Nguyen and Nguyen 2020). The findings further revealed that English plays a significant role in Vietnam, influencing sociocultural aspects, teaching materials, and even lecturers'

perceptions of their English teaching skills and academic subject instruction.

A critical issue identified in the research was the absence of cultural representation from non-native English varieties in teaching materials. This lack of diversity may lead to increased resistance and confusion among learners when interacting with speakers from different linguistic and cultural backgrounds. Without exposure to varied English accents and discourse styles, students may develop a limited understanding of English as a global language, reinforcing monolithic language ideologies (Jenkins 2015). Moreover, lecturers reported having no authority over the selection of teaching materials, as these decisions were made at the institutional level. Consistent with Walsh (2010), the research found that teachers were required to follow textbooks mandated by the board of management, limiting their ability to incorporate diverse linguistic and cultural elements into their instruction. Addressing these issues requires institutional reforms that promote lecturers' autonomy and a more inclusive approach to English language teaching in EMI classrooms.

The research found that lecturers had no control over the selection of teaching materials, as these decisions were determined at the institutional level rather than by individual instructors. This finding is consistent with Walsh (2010), who reported that lecturers lacked the autonomy to choose their instructional resources, resulting in students being primarily exposed to native English varieties through pre-recorded tapes. Similarly, interview participants in this study stated that when working in schools, universities, or English centers, they had no authority over textbook selection, as this responsibility fell solely on the board of management. Lecturers were required to follow the prescribed curriculum without any input, limiting their ability to adapt materials to better suit their students' linguistic backgrounds and learning needs. This institutionalized approach reinforces the dominance of native-speaker English norms in EMI classrooms and restricts lecturers from incorporating diverse English varieties into their instruction.

5.2 Limitations and Recommendations

Beyond acknowledging the research's limitations, several recommendations for further research on English teaching practices in EMI classrooms can be proposed. First, since Vietnamese lecturers' perceptions of EMI settings and policies are shaped by interactions with students, school administrators, and policy-makers, future research should broaden their scope to include these key stakeholders. Investigating the perspectives of learners, managers, and institutional decision-makers would provide a more comprehensive understanding of how EMI is implemented and perceived in Vietnam. This approach would align with previous research highlighting the influence of institutional policies on lecturers' pedagogical choices and attitudes toward English language instruction (Macaro and Searle 2025; Walsh 2010).

Second, Honna and Takeshita (2014) suggested that incorporating metaphor-based learning strategies (using metaphors to connect new concepts with familiar experiences) can enhance the understanding of English as a language that embodies multiple cultures. Researchers could explore this method further by designing experimental studies that introduce metaphor-based learning in EMI classrooms. By measuring the shifts in teachers' and students' perceptions before and after such interventions, scholars could gain deeper insights into how EMI practices shape attitudes toward English as a global language. Additionally, such research could help challenge the dominance of native-speaker norms by promoting an appreciation for linguistic diversity in English education.

Finally, because this research may lack generalizability due to its sample size and research scope, further qualitative studies should include a broader range of participants across different educational levels and regions in Vietnam. Expanding the research population would provide a more nuanced understanding of EMI teaching practices and their challenges. More extensive studies would also contribute to raising awareness among Vietnamese educators and policy-makers, fostering discussions about the need for inclusive and adaptable English teaching practices in EMI settings.

REFERENCES

- Biesta, Gert and Tedder, Michael. 2006. How Is Agency Possible? Towards an Ecological Understanding of Agency-as-Achievement. *Learning Lives: Learning, Identity, and Agency in the Life Course*. Teaching and Learning Research Programme.
- Borg, Simon. 2006. *Teacher cognition and language education: Research and practice*. Continuum.
- Bourdieu, Pierre. 1991. *Language and symbolic power*. Polity Press.
- Creswell, John W. and Creswell, John David. 2018. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE.
- Dang, Thi Kim Anh and Cross, Russell. 2023. Globalizing Teacher Education Through English as a Medium of Instruction: A Vygotskian Perspective. In: Menter, I. (Eds), *The Palgrave Handbook of Teacher Education Research*. Palgrave Macmillan, Cham.
- Dearden, Julie. 2015. *English as a medium of instruction: A growing global phenomenon*. British Council.
- Dörnyei, Zoltán. 2007. *Research Methods in Applied Linguistics: Quantitative, Qualitative, and Mixed Methodologies*. Oxford University Press.
- Honna, Nobuyuki and Takeshita, Yuko. 2014. *The Pedagogy of English as an International Language*. Springer Nature.
- Jenkins, Jennifer. 2015. *Global Englishes. A Resource Book for Students*. Routledge.
- Macaro, Ernesto. 2018. *English Medium Instruction: Content and Language in Policy and Practice*. Oxford University Press.
- Macaro, Ernesto and Aizawa, Ikuya. 2022. Who owns English medium instruction? *Journal of Multilingual and Multicultural Development*, 1-14.
- Macaro, Ernesto and Searle, Mark. 2025. *Navigating English Medium Instruction: A Student Handbook*. Routledge.
- Malmström, Hanna and Zhou, Shaofu. 2025. *Language–Subject Teacher Collaboration in English-Medium Higher Education: Current Practices and Future Possibilities*. RELC Journal.
- Nguyen, Huong Thu. 2018. English-medium instruction management: The missing piece in the internationalisation puzzle of Vietnamese higher education. In Ly Thi Tran and Simon Marginson (Eds.), *Internationalisation in Vietnamese higher education* (pp.119-137). Springer.
- Nguyen, Huong Thu. 2022. English Medium Instruction in Vietnamese Higher Education: From Government Policies to Institutional Practices. *English Medium Instruction Practices in Higher Education*. London, Bloomsbury Academic. 173-184.
- Nguyen, Thi Mai Hoa. 2011. Primary English language education policy in Vietnam: Insights from implementation. *Current Issues in Language Planning*, 12:225-249.
- Nguyen, Huong Thu, Walkinshaw, Ian, and Pham, Hiep Hoa. 2017. EMI Programs in a Vietnamese University: Language, Pedagogy and Policy Issues. In: Fenton-Smith, Ben, Humphreys, Pamela, and Walkinshaw, Ian. (Eds.), *English Medium Instruction in Higher Education in Asia-Pacific* (pp.37-52). *Multilingual Education*, 21. Springer.

- Nguyen, Thi Phuong Doan and Nguyen, Van Loi. 2020. EMI in Vietnam: What High School Teachers Think and Do. *International Journal of Language Teaching and Education*, 4:36-52.
- Nguyen, Van Huy, Tran, Phuong Thi Thao, and Ngo, Phuong Le Hoang. 2024. English as a Medium of Instruction in Vietnamese Universities: A Perspective of Language Policy and Planning. In Hung, Bui, Truong, Bang, and C. Nguyen (Eds.), *Teacher and Student Perspectives on Bilingual and Multilingual Education* (pp.98-113). IGI Global Scientific Publishing.
- Pennycook, Alastair. 2017. *The Cultural Politics of English as an International Language*. Taylor and Francis.
- Pham, Minh, and Doan, Ba Ngoc. 2020. English as a medium of instruction (EMI) in Vietnamese universities: Policies of encouragement and pedagogies of assumption. In Ha Le Phan and Ngoc Ba Doan (Eds.), *Higher education in market-oriented socialist Vietnam: New players, discourses, and practices* (pp.261-282). Springer.
- Pham, Min, Nguyen, Viet Nga, Hoang, Hoa, Nguyen, Thi Thien Huong, Le, Duc Hanh, and Barnett, Jenny. 2023. Internationalisation and English medium education in Vietnamese universities: Professional role transformation among English language lecturers. *Research in Comparative and International Education*, 18:292-316.
- Priestley, Mark, Biesta, Gert, and Robinson, Sarah. 2015. *Teacher Agency: An Ecological Approach*. Routledge.
- Ricento, Thomas and Hornberger, Nancy. 1996. Unpeeling the Onion: Language Planning and Policy and the ELT Profession. *TESOL Quarterly*, 30:401-427.
- Rose, Heath, Curle, Samantha, Aizawa, Ikuya, and Thompson, Gene. 2020. What drives success in English medium taught courses? The interplay between language proficiency, academic skills, and motivation. *Studies in Higher Education*, 45:2149-2161.
- Sahan, Kari, and Şahan, Özgür. 2022. Content and Language in EMI Assessment Practices: Challenges and Beliefs at an Engineering Faculty in Turkey. In: Kirkgöz, Y., Karakaş, A. (Eds), *English as the Medium of Instruction in Turkish Higher Education. Multilingual Education*, 40. Springer.
- Spolsky, Bernard. 1998. *Sociolinguistics. Oxford introductions to language study*. Oxford University Press.
- Tollefson, James, and Tsui, Amy. 2014. Language diversity and language policy in educational access and equity. *Review of Research in Education*, 38:189-214.
- Tran, Ly Thi, Ngo, Nga Thi Hang, Nguyen, Hoa Thi Mai, and Le, Truc Thi Thanh. 2022. Labour Market Demands, Graduate Employability and Employment Outcomes in Vietnam. In *Employability in Context*. Springer.
- Tran, Ly Thi, and Nguyen, Huong Thu. 2018. Internationalisation of higher education in Vietnam through English medium instruction (EMI): Practices, tensions and implications for local language policies. In I. Liyanage (Eds.), *Multilingual education yearbook: Internationalization, stakeholders and multilingual education contexts* (pp.91-106). Springer.

- Tran, Phat, Tran, Thao, and Le, Nhung. 2026. Exploring lecturers' perceptions about teaching practices in EMI classrooms. *Proceedings of International Academia*, 1:1-10.
- Tri, Dang, and Moskovsky, Christo. 2023. Language use in English-medium instruction programs in Vietnamese higher education: From policy to practice. *Asian Englishes*, 25:326-342.
- Vu, Nha Thi Thanh, and Burns, Anne. 2014. English as a Medium of Instruction: Challenges for Vietnamese Tertiary Lecturers. *TESOL Quarterly*, 48:813-835.
- Walsh, Maureen. 2010. Multimodal Literacy: What Does It Mean for Classroom Practice? *Australian Journal of Language and Literacy*, 33:211-239.
- Zheng, Qun, and Choi, Tae-Hee. 2024. English-medium instruction as an internationalisation strategy at a second-tier Chinese university: instructors' challenges and their shaping factors. *Asian-Pacific Journal of Second and Foreign Language Education*.

Tran Tan Phat
Faculty of Foreign Languages
Ho Chi Minh University of Banking
56 Hoang Dieu II street, Thu Duc ward, HCMC, Vietnam
phattt@hub.edu.vn

越南英語授課講師項目對於國家語言政策的看法

Tran Tan Phat

胡志明市銀行大學

胡志明市開放大學

本研究探討了教師在越南英語授課(EMI)專案中對於國家語言政策的看法，重點在於這些看法要如何塑造課堂教學實踐。該研究採用定性研究設計，在胡志明市銀行大學對於來自不同學科的十位 EMI 教師進行了研究。同時，採用了問卷調查、課堂觀察和訪談收集資料對全面教師的既定觀念與實際教學實踐進行研究。研究結果顯示教師對 EMI (英語授課) 獲取了有效的國際知識，提升了學生就業的能力。本研究跟著教師們的反映，在實施 EMI 政策的過程中提出了多種挑戰性，包括工作量增加、難以平衡的專業內容與語言教學障礙，以及缺乏明確的政策指導方式進行研究並找出解決方案。

關鍵字：EMI 政策；教學實踐；教學障礙

版權頁資訊

Current Issues in L2 Teaching and Research, Vol. 2 [當代二語教學與研究第二冊]

主編順序

Stano Kong [江丕賢], Miao-Chin Chiu [邱妙津], Yu-Hsin Huang [黃郁欣], Kai-Lin Wu [吳凱琳], Pin-Hua Wu [吳品嬋], Yu-Han Lin [林鈺涵], One-Soon Her [何萬順]

編輯助理

Lawrance A. Wu [吳裕仁], Lucas Tony Wilds [郭洛德], Toyama Kokori [當山心梨]

Current Issues in L2 Teaching and Research, Vol. 2

作者：Stano Kong (江丕賢)、Miao-Chin Chiu (邱妙津)、Yu-Hsin Huang (黃郁欣)、
Kai-Lin Wu (吳凱琳)、Pin-Hua Wu (吳品嫻)、Yu-Han Li (林鈺涵)、
One-Soon Her (何萬順)

發行人：張富恭

出版者：文鶴出版有限公司

地址：106 台北市金山南路二段 200 號 8 樓

電話：(02) 2393-4497

傳真：(02) 2394-6822

總公司：106 台北市金山南路二段 200 號 8 樓

TEL: (02) 2393-4497 FAX: (02) 2394-6822

各區業務及門市

北區：106 台北市金山南路二段 200 號 8 樓

TEL: (02) 2393-4497 FAX: (02) 2394-3810

南區：802 高雄市同慶路 88 號 2 樓

TEL: (07) 2270-888 FAX: (07) 2270-801

網路書店：<http://www.crane.com.tw>

法律顧問：蘇家宏律師 TEL: (02) 2322-3588

劃撥帳號：01079261

戶名：文鶴出版有限公司

出版日期：2026 年 08 月初版一刷

I S B N：978-626-7886-19-9

書號：0031588

※ 著作權所有，翻印必究，非經本公司同意不得轉載、仿製或局部使用於雜誌、網路或其他書籍、媒體上。

※ 若有缺頁，裝訂錯誤，煩請寄回更換。